Prognosis of Chronic Renal Syndrome by Classification and Progression Using Temporal Abstraction

Priya Sampath, Kannan Arputharaj

Abstract: Chronic renal syndrome is defined as a progressive loss of renal function over period. Analysers have make effort in attempting to diagnosis the risk factors that may affect the retrogression of chronic renal syndrome. The motivation of this project helps to develop a prediction model for level 4 CKD patients to detect on condition that, their estimated Glomerular Filtration Rate (eGFR) stage downscale to lower than 15 ml/min/1.73 m². End phase renal disease, after six months accumulating their concluding lab test observation by assessing time affiliated aspects. Data mining algorithm along with Temporal Abstraction (TA) are confederated to reinforce CKD evolvement of prognostication models. In this work a inclusive of 112 chronic renal disease patients are composed from April 1952 to September 2011 which were extracted from the patient's Electronic Medical Records (EMR). The information of chronic renal patients are collected in a big spatial info-graphic data.

In order to analyse these info-graphic data, it is significant to detect the issues affecting CKD deterioration and hence it becomes a challenging task. To overcome this challenge, time series graph has been generated in this project work based on creatinine and albumin lab test values and reports of the time period. The presence of CKD diagnostic codes are transformed into default seven digit default format of International Classification of Disease 10 Clinical Modification (ICD 10 CM). Feature selection is performed in this work based on wrapper method using genetic algorithm. It is helpful for finding the most relevant variables for a predictive model. High Utility Sequential Rule Miner (HUSRM) is used here to address the discovery of CKD sequential rules based on sequence patterns.

Temporal Abstraction (TA) techniques namely basic TA and complex TA are used in this work to analyse the status of chronic renal syndrome patients. Classification and Regression Technique (CART) along with Adaptive Boosting (AdaBoost) and Support Vector Machine Boosting (SVMBoost) are applied to develop the CKD in which the progression prediction models exhibit most accurate prediction. The results obtained from this work divulged that comprehending temporal observation forward the prognostic instances has escalated the efficacy of the instances. Finally, an evaluation metrics namely accuracy, sensitivity, specificity, positive likelihood, negative likelihood and Area Under the Curve (AUC) are helps to evaluate the performance of the prediction models which are designed and implemented in this project.

Key Words: CKD, progression, time series data, genetic algorithm, sequential rules, TA classification and prediction model.

I. INTRODUCTION

Chronic renal syndrome is a disease being that patients are deliberately reduce the behaviour of renal. Chronic renal disease has evoked significant consideration in the social

healthiness concern in current decades. Almost, probability of 10 percent of global population agonise of this syndrome [1]. Chronic renal disease affects abundant illness along with notably affects the quality of patient's life; consequently, health professionals from global wide are undertaking to acquire the aid of averting its occurrence [2, 3].The initial ailment of chronic renal disease are generally wide-ranging, there are same to the particular some various ailments. Chronic renal syndrome is typically not detected up to the time of additional crucial ailments including edema, haematuria and blood clots occurs [4]. At that moment, renal performance decreased 5% below normal, the body cannot manage regular metabolism then the patient initialize last phase kidney syndrome. In order to extend the living span, patient's last phase renal syndrome should require prolong with the help of dialysis or a kidney transplantation [5, 6]. A report by the National Kidney Function (2015) revealed 10% of world population are affected by CKD. The study of Global Burden of Disease (2010), chronic renal disease was ranked 27th overall the list of major effects of increase number of deaths globally in 1990, but it rose to 18th rank in the year 2010. Of the 2 million of global population currently undergoing treatment with kidney transplant or a dialysis to stay alive, but only 10% of population who actually need treatment to save their lives. Over 2 million people receive treatment for renal failure, only the five countries are majority treated they are United States, Japan, Germany, Italy and Taiwan. In the country USA, patients management and their doctoring costs regarding ESRD nearly $26.8 billion and hospitalisation of CKD is likely to exceed $48 billion in a year. Medication treatment for renal failure consumes 6.7% of the overall Medicare budget in that, they care for less than 1% of the total population. [7]. According to Taiwan Society of Nephrology (2009), currently 60,000 people undergoing regular treatment for ESRD with increasing of 8000 people in each year [8]. Healthcare envelopes the explained the process of prognosis, medicaments and impediment of syndrome, trauma and some other mental and physical deterioration in humans.

The combination of data mining technique and healthcare application have developed as some of reliable primordial detection systems and healthcare related systems from the clinical and diagnosis data. Time series graphs are mandatory in variety of applications of statistics. When reporting the values of same variable over an elongated period of time. Consistently, it is arduous to discern any trend or pattern.

Sequential rule mining is practiced to a common database in order to discover rules which involves a patient's age, sex and complete medical history. On comprehending the rules into present medical care, a patient should be featured as susceptive to the future most ailments based on either past or present ailment conditions, patient's gender and their birth of year.

Temporal Data Mining (TDM) referred as the exercise of focusing on either interesting patterns or correlations in huge temporal datasets. TDM has developed from data mining and was increasingly manipulated in the fields of temporal reasoning and temporal databases. In data mining, Classification modelling function which allocates items in terms of group to target classes or categories. The main aim for classification is to exactly anticipate to find the targeting class for respective cases in the data. Classifiable sequence patterns are generated and which identifies whether he/she is a chronic kidney patient or not.

The continuation on this subject is coordinated in such a way. In the background literature survey portion. Subsequently, define with the mining issues, chronicles of sufferers with renal syndrome and then elucidate with the progression models. Then it follows, experimental results are presented. Finally, presents the conclusions.

II. PROBLEM STATEMENT

In recent days, Chronic Kidney Diseases are the major killers in the modern era. Patient's having CKD does not exhibit symptoms until stage 3 or 4 CKD. Early prognosis of severity and progression of chronic kidney syndrome is a challenging task. Because of the increasing burden of last phase renal disease i.e. (ESRD), it has better to assess the disease at the earliest.

III. LITERATURE SURVEY

The techniques used in the literature for the patients personal and medical dataset collection, feature selection, sequential risk rules mining, Temporal Abstraction, classification modelling and chronic kidney disease prediction are discussed in this chapter.

A. Dataset Collection

Dataset collection plays a vital role in early assessment of chronic kidney disease. Many methods are available in the literature for data collection. Among them, Li Chen Cheng et al. [9] analysed a sole of the biggest renal centre in a cosmopolitan medical centre in Taiwan using a new technique. The records in this database was documented in International Classification of Diseases and they adopted the data format of the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9 CM). These database includes prognostic details, patient's hospital admissions, doctor's prescriptions, patient's disease history profiles, etc., and the prognostic details are recorded with every three ICD-9 codes at almost in every patient's visit. Khotimah et al. [10] defined with a constant period of data with the set of approximately 53,444 patinet's records were generating origin from description of detailed data for almost 2,816 type 2 diabetes patients details presented by Kyoto University Hospital. Whereas, the data set which includes the information detailed about the patient id (pid), patient's hospitalisation start time, their end time, and the combining of medicament categories. Where the patients personal and medical records dataset are collected from Electronic Medical Records (EMR) database [11]. The records collected in EMR database adopts the data format of the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10 CM) [12].

B. Feature Selection

H Polat et al. [13] suggested a recent feature selection method to change the class distribution through adding virtual samples, which are generated by filter and wrapper method. Their proposed method analyse the additive effects in addition to examine the multiplicative effect between samples. In this method, the use of filter method by calculating information gain and then calculate symmetric uncertainty. In contrast to wrapper method the authors used genetic algorithm for optimization and then calculate fitness value for each attributes and performed feature selection using filter method. The main drawbacks of filter method is that ignores the interaction with classifier whereas in wrapper, the risk of over fitting is present and it is computationally intensive.

C. Mining Sequential Risk Rules

Sujeevan Aseervatham et al. [14] proposed bitSPADE, a novel algorithm that coalitions the top aspect of Sequential PAttern Mining (SPAM), this is one of the speediest algorithm, and SPADE is one among the high memory sufficient algorithm and recent pruning approach which allows bitSPADE to attain the highest performances. Sequential PAttern Discovery using Equivalence classes (SPADE) is a candidate generation and then test algorithm. The main drawback of bitSPADE algorithm is that the performance in real life applications is to abstract the sequential patterns under particular period of time constraints is less.

Souleymane Zida et al. [15] proposed a novel algorithm that coalitions various maximisation to mining huge efficacy consecutive ethics and determining conception bringing forth a great acquisitions in database. Frequent Pattern Mining (FPM) is used to discovering the collections of effects looking collected repeatedly in the transaction data collection. High Utility Sequential Rule Miner (HUSRM) discovers pattern based on calculation of utility values. The main drawback of this algorithm is less space efficient and memory usage is comparatively high where compression cannot be included.

54

D. Temporal Abstraction

Li Chen Cheng et al. [1] proposed Temporal Abstraction (TA) deal with the extrication of subjective traits of time sequence derived from the rules which are evident by medical expertise. Temporal Abstraction (TA) detailed with various kinds of info-graphic form in time series data, there are statuses, trends and some other complex time correlated features. In most cases, there are two types of Temporal Abstraction (TA). They are complex TA and basic TA. Complex Temporal Abstraction determine the interdependent with temporal series of interval. Whereas, Basic Temporal Abstraction is defined with representational time series data or detection of numerical data and then section which are performed by qualitative method. There are two kinds of Temporal Abstraction parameter obtained where, state Temporal Abstraction parameter that defect the patterns of qualitative functioning which corresponds to high, low and normal values and trend Temporal Abstraction parameters include decreasing, increasing, and constant figure in a time serial numerical value.

E. Classification Modeling Chronic Kidney Disease Prediction

Chu Yu Chin et al. [16] determined earlier risk patterns communicate disease with a target condition using data mining technique by evaluating Rheumatoid Arthritis (RA) premorbidity for detecting threat patterns in wide-ranging real data. The primary approach involves risk pattern mining using Associative Classification technique. The major drawback of this technique is imbalance of data issue because of insufficient time interval extrication. The key challenges are class imbalance issue, mining sequential rules, Temporal Abstraction and classification based on accuracy and time interval extraction for the prediction of chronic renal disease. Genetic algorithm to resolve the class imbalance problem, High Utility Sequential Rule Miner (HUSRM) algorithm to identify sequential risk rules based on utility values, Temporal Abstraction (TA) for generting rules based on season wise analysis and Adaptive Boosting (AdaBoost) and Classification and Regression Technique (CART) algorithm classification in order to get high accuracy are used to overcome the above mentioned challenges.

IV. SYSTEM DESIGN

The design of the system is to build the sequential and persistent classification modelling for chronic renal syndrome prognosis. The steps in the design process includes model building and model evaluation stage. The model building stage includes data pre-processing, sequential rule generation, Temporal Abstraction and classification modelling. The model evaluation stage includes evaluation of training and testing set by generating a line graph and it also includes the calculation of metrics such as sensitivity, specificity, positive likelihood, negative likelihood and accuracy.

The dataset consists of 112 CKD patients personal and medical records which serves as the input. The dataset also comprises of ICD 10 CM (international Classification of Diseases Tenth Revision Clinical Modification) diagnostic codes. ICD 10 CM gives more clinical details and updated information about chronic diseases. The first step in the proposed framework is data preprocessing. The two phases of Data Preprocessing are data transformation (inconsistent noise removal) and feature selection (solving imbalance issues).

The next step is generating sequential rules using High Utility sequential Rule Miner (HUSRM) algorithm. The sequential risk patterns identified by calculating utility values. Based on the sequential rules are generated by support and confidence ratios are calculated. Temporal Abstraction (TA) used to analyse the season wise analysis in basic TA and complex TA. The overall architecture is depicted in Figure 1.

System Architecture
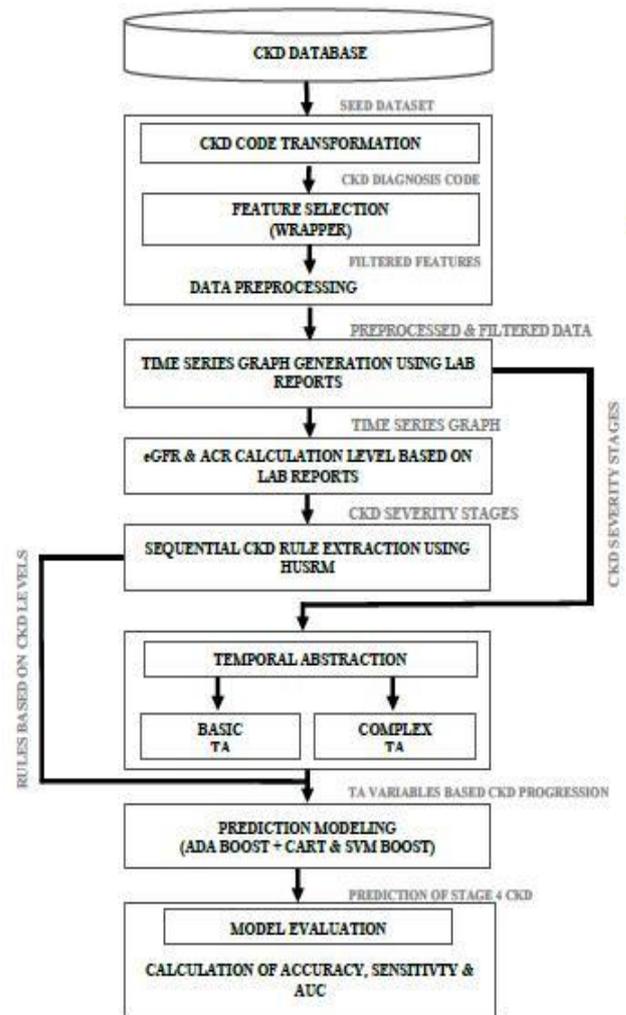
Figure 1: Chronic Renal Syndrome Prognosis Architecture

The third step deals with modelling the classifiable for identifying stage wise chronic renal disease. To improve the accuracy combination of AdaBoost and CART are used in this work along with SVMBoost algorithm.

55

The final step is post analysis which consists of two stages namely, time interval extraction and novelty measuring. Based on the result of classification modelling and post analysis early assessment of last phase renal disease and prognosis of chronic renal diseases has been carried out effectively.

The evaluation stage consists of validating the testing and training dataset use of K-fold cross validation technique and the calculating sensitivity, specificity, positive likelihood, negative likelihood and accuracy. A line graph is generated which indicates the sensitivity, specificity and accuracy values by taking into account performance measures and minimum support threshold. A bar graph is generated which indicates the different age groups of patients who are affected by stage wise chronic renal disease.

V.   MODULES DESCRIPTION

The System architecture highlights the following functional modules.
- A.   Data Collection and Feature Selection
- B.   Time Series Graph Generation and eGFR, ACR Calculation
- C.   Sequential Rules Generation
- D.   Temporal Abstraction
- E.   Classification Modelling
- F.   Model Evaluation

The following subsection provides a detailed view of the functional modules.

A.   Data Collection And Feature Selection

Dataset details and pre-processing steps involves patients and personal medical records includes the attributes such as patient id, admission id, admission start and end date, primary diagnosis code and description, lab name, value, units, date and time respectively. It consists of 112 patients personal and medical records in which lab reports of 32,967 records serves as a input. The dataset also comprises of International Classification of Diseases Clinical Modification (ICD 10 CM) diagnostic codes. ICD 10 CM gives more clinical details and updated information about chronic diseases. Lab reports of 32,967 and patient reports of 112 records consists of patient's race, marital status, country and age respectively. There are two steps involved in data collection and pre-processing are:

In the first step, inconsistent noise removal where the data transformation of diagnostic diseases code into seven digits default data format to avoid coding errors. It examines the prognostic detailed records and eject the prognostic International Classification of Diseases Tenth Revision Clinical Modification (ICD 10 CM) codes with coding errors. Furthermore, it extracts the patient's visiting date and prognostic detailed medical records are defined in a sequential order. Additional transformation of International Classification of Diseases Tenth Revision Clinical Modification (ICD 10 CM) codes into five digits form and they will use the format as the default data form. For example, ICD-10 code N18.1 and N18.5 are transformed to N181000 and N185000 respectively.

In the next step feature selection, wrapper method using genetic algorithm to calculate fitness value for each features by evaluating information gain values using filter method thereby calculating symmetric uncertainty for each attribute.

B.   Time Series Graph Generation and eGFR, ACR Calculations

Time related laboratory test information are extracted from the patient's laboratory time series data. In that laboratory test information carried out the number of occurrences of creatinine and albumin for each patient clinical history that are cumulate in large database of Electronic Medical Records (EMR) about to particular period, that tends time series in high dimensional data into high dimensional time series graph. Where, X axis represents DD-MM-YY (Lab test taken by the patient's) and Y axis represents Creatinine mg/dl and Albumin gm/dl has been generated for 112 chronic renal patients.

Estimated Glomerular Filtration Rate (eGFR) calculation of patient's are mainly depends on lab units. Estimated Glomerular Filtration Rate (eGFR) is the best experiment to calculate the phase of renal functioning and have an impact on the phase of kidney syndrome. If eGFR level less than 15 ml/min/1.73m then patient ends up with End Stage Renal Disease (ESRD). For determining CKD stages based on eGFR rates are ranges at below:

Stage 1 → > 90
Stage 2 → 60 to 89
Stage 3 → 30 to 59
Stage 4 → 15 to 29
Stage 5 → < 15

Albumin to Creatinine Ratio (ACR) is evaluated by dividing the ratio of albumin concentration in milligrams by creatinine concentration in grams.

Normal to Mild Increase   → ACR < 3
Moderately Increase   → ACR 3 to 30
Severely Increases   → ACR > 30

Time Series Graph Generation

The time series graph has been generated based on albumin and creatinine values of each patient are shown below in Figure 2.
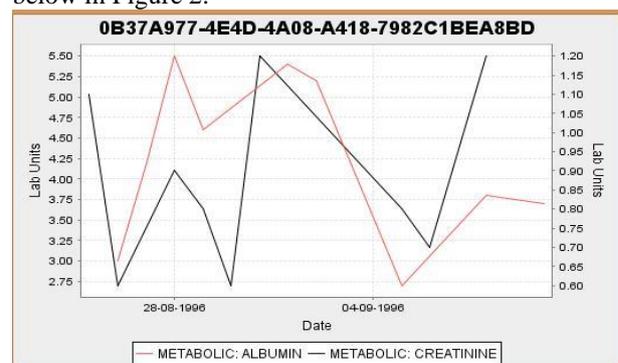
Figure 2: Time Series Graph

C. Sequential Rule Generation

The frequent sequences are generated and are used to produce rules.

A utility value is calculated for every 112 CKD patients. A rule R: (X → Y) may be performed and accompanied by Sup(X) and Conf(R), where Conf(R) is the Confidence of the sequential rule and is represented as below in formula.

Conf(X → Y) = Sup(X→Y) /Sup(X) After finding out support, confidence and utility fix a minimum threshold support and minimum threshold confidence then generate a CKD severity stages for each and every lab names based on high utility values of CKD severity stage has mentioned for each patients.

D. Temporal Abstraction

By applying Temporal Abstraction, to obtain

features based on the different phase of patient who are affecting with renal syndrome, whereas, temporal abstraction concepts were applied to extract the time related parameters based on time series graph has been generated for CKD lab names with season wise analysis.

Basic TA mechanism extracts trend changes from time series. Where trend TA represents increasing, stationary and decreasing patterns. Then, categorize the variable in the datasets. Extract the lab units of patients and fix the threshold, then eliminate the lab units less than the threshold from that identify the status. The status are represents in a seven possible state, there are:

XH → Extremely High
H → High
N/H → Higher than Normal
N → Normal
N/L → Lower than Normal
L → Low
XL → Extremely Low

Whereas Complex TA represents with the season wise patterns where CKD lab values are calculated by average value of two adjacent item for that both state TA represents high, normal and low are generated and then for trend TA increasing, stationary and decreasing patterns are extracted based on the status. TA Input And Output: The input of temporal abstraction based on time series graph has been generated for each patient with their lab tests based on season wise analysis are as shown in Figure 3.
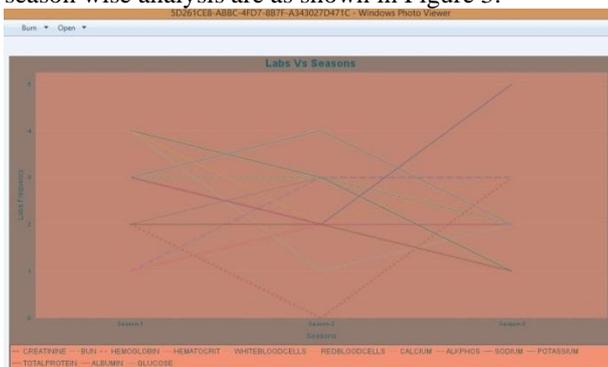
Figure 3: Season Wise Time Series Graph

E. Classification Modeling

By applying classification of AdaBoost and CART technique algorithm, the results of stage wise chronic kidney patients are obtained. This is achieved by taking the high

utility sequential Rule Miner output which match with frequent item rules from temporal abstraction algorithms output then the chronic renal patient has been categorized into stage wise CKD with respect to their respective patients id.

F. Model Evaluation

Sensitivity, specificity, positive likelihood, negative likelihood and accuracy are used to calculate the exact rate in every class, although accuracy is a measure for assessing the evaluation of prediction model. Finally, a line graph is generated which indicates the sensitivity, specificity and accuracy values by taking into account performance measures and minimum support threshold of patients who are affected by chronic kidney disease.

VI. DISCUSSION

In this, Phase 4 chronic renal syndrome patients has poor comprehension of their syndrome condition, that indicates their hospitalisation should improve patient's care and giving awareness about the knowledge of health at every phase. Patients must attain up-to some height of consciousness then insightfully should develop their one's health care that helps to guide them to reins the syndrome and attenuate the worsening of renal functionality. Informational courses should provide to the public regarding proper medicine guidance, diet plan, regular physical exercise and treatment of dialysis are also be advised. During the final stage, most of the patients undergoes to dialysis treatment and a transplantation of kidney should be done where in the case of not attention to be paid towards the syndrome progression. To the extent that, AdaBoost (AdaBoost) and Classification And Regression Technique (CART) model enacted the superior enactment along with all prediction models. The output from HUSRM and TA module, results of every rule in the level 4 CKD progression where patients ends with End Stage Renal Disease (ESRD) later 6 months. Temporal Abstraction based parameters endeavour the critical consequences, about all significant rule comprises complex Temporal Abstraction concepts, referring the long lasting following of variations in lab experimental results can confer a prior prediction and expedite the approach of attenuate the degradation of chronic renal syndrome. For antiquated patients with the level 4 chronic renal syndrome, declination of BUN and Creatinine results are mandatory in the lab results. Moreover, when the patient has a deficient habitual, there are inactivity, without physical exercise, smoking, and some other complications that includes some other ailments includes, cardiac syndrome or diabetics tends to last phase renal diseases that can occurs readily.

VII. CONCLUSION AND FUTURE WORK

Analysing the system using evaluation metrics such as sensitivity, specificity, positive likelihood, negative likelihood and accuracy can be used for further improving the results obtained from the designed system. It lists the shortcomings and possible extensions for future work.

A. *Conclusion*

The system developed in this project work bridges the knowledge of medical conception by performing a novel model on discovering sequential patterns for status of chronic renal disease. Meanwhile, the discovered patterns based on high dimensional time series data by applying temporal abstraction convey the beneficial potential for patient medical researchers to build furthermore idealised examination by discovering trend markers and better treatments of mandatory seriousness of chronic renal disease status such as CKD severe, end stage renal diseases.

Therefore, the sequential pattern mining using classification techniques helped to provide effective decision based on chronic kidney syndrome. Hence, risk prediction modelling is more efficient in assessing the chronic kidney disease at the early stage and disease progression modelling is used in this project work to measure the progress of the end stage chronic renal disease and control it.

B. Future Work

Future work includes further analysis and improvements in the existing system to identify the early assessment of last stage chronic kidney syndrome effectively by applying molecular biomarkers that would aid the diagnosis, assessment and guidance of theory of patients with chronic renal disease.

REFERENCES

1. Chen, N. A. N., HSU, C. C., Yamagata, K., & Langham, R. (2010). Challenging chronic kidney disease: experience from chronic kidney disease prevention programs in Shanghai, Japan, Taiwan and Australia. *Nephrology*, *15*(s2), 31-36.
2. National Kidney Foundation, about chronic kidney disease. Available: http://www.kidney.org/kidneydisease/aboutckd.cfm, 2015.
3. Thomas, R., Kanso, A., & Sedor, J. R. (2008). Chronic kidney disease and its complications. *Primary care: Clinics in office practice*, *35*(2), 329-344.
4. Kronenberg, F. (2009). Emerging risk factors and markers of chronic kidney disease progression. Nature Reviews Nephrology, 5(12), 677.
5. Chen, Y. S., & Cheng, C. H. (2013). Application of rough set classifiers for determining hemodialysis adequacy in ESRD patients. *Knowledge and information systems*, *34*(2), 453-482
6. Abbasi, M. A., Chertow, G. M., & Hall, Y. N. (2010). End-stage renal disease. *BMJ clinical evidence*, *2010*.
7. https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease
8. Yang, W. C., Hwang, S. J., & Taiwan Society of Nephrology. (2008). Incidence, prevalence and mortality trends of dialysis end-stage renal disease in Taiwan from 1990 to 2001: the impact of national health insurance. *Nephrology dialysis transplantation*, *23*(12), 3977-3982.
9. Cheng L C, Hu Y H and Chiou S H. Applying the Temporal Abstraction Technique to the Prediction of Chronic Kidney Disease Progression. IEEE Journal of medical systems, Vol 41, No. 5, pp. 85, 2017.
10. Khotimah P H, Yoshikawa M, Hamasaki A, Sugiyama O, Okamoto K and Kuroda T. Comparing frequent patterns: A study case of Apriori and singleton implementations in a diabetes type 2 data set. In Computer, Control, Informatics and its Applications (IC3INA), pp. 163-168, 2016.
11. Dataset details: http://www.emrbots.org/Downloads2222694716.html.
12. Details of International Classification of Diseases Tenth Revision Clinical Modification (ICD 10 CM):http://www.icd10data.com/ICD10CM/Codes.
13. Polat H, Mehr H D, and Cetin A. Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. Journal of medical systems, Vol 41, No. 4, p. 55, 2017.
14. Aseervatham S, Osmani A and Viennet E. bitSPADE: A lattice-based sequential pattern mining algorithm using bitmap representation. Sixth International Conference on IEEE., pp. 792-797, 2006.
15. Zida S, Fournier Viger P, Wu C W, Lin J C W, and Tseng V S. Efficient mining of high-utility sequential rules. n International Workshop on Machine Learning and Data Mining in Pattern Recognition, pp. 157-171.
16. Chin C Y, Weng M Y, Lin T C, Cheng S Y, Yang H K and Tseng V S. Mining Disease Risk Patterns from Nationwide Clinical Databases for the Assessment of Early Rheumatoid Arthritis Risk. PloS one, Vol 10, No. 4, 2015.

AUTHORS PROFILE

First Author Priya Sampath, MTech (Information and Technology), 2016 – 2018, College of Engineering Guindy, Anna University, Chennai – 600025. sampathpriyaa@gmail.com

Second Author Kannan Arputharaj, VIT Vellore, Retired Professor of Anna University, Chennai – 600025. kannan@annauniv.edu