International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 20 (2017) pp. 9500-9505

A Prediction of Chronic Kidney Disease Using Feature based Priority Assigning Algorithm

S.Dilli Arasu
Research Scholar, Department of Computer Applications (Ph.D.),
Bharath University, # 173 Agharam Road, Selaiyur, Chennai - 600 073,Tamil Nadu, India.
Orcid Id: 0000-0002-4961-8130

Dr. R.Thirumalaiselvi
Research Supervisor / Assistant Professor, Department of Computer Science,
Government Arts College for Men (Autonomous)- Nandanam,
Chennai -600 035, Tamil Nadu, India.

Abstract

Kidney disease is one of most dangerous disease spreading all over the world due to the change in our life style, including food habits, environment, etc. so it is essential to choose any remedy to avoid as well as predict the disease in early stage which helps to save the life. The prediction of kidney disease can be done efficiently using Priority assigning algorithm. It is done with the help of Attribute selection method. Attribute (Feature Selection) selection to support kidney disease detection. It is made by eliminating the attributes of less importance and also to select the important attribute to be present in the dataset. But existing algorithm does not suit for imbalanced dataset or large number of distinct values. In this paper, a new algorithm called priority assigning algorithm has been introduced in order to assigns high priority for more important features. This makes the classification process efficient and kidney disease can be predicted efficiently.

Keywords: SVM, ANN, PLS-DA, MFPT, WAELI, WAELI-FPA, Data Mining.

INTRODUCTION

Early detection and treatment can even often evade chronic kidney disease from getting bad. Evolution of kidney disease, may ultimately lead to kidney failure, which requires dialysis or a kidney transplantation to sustain life [1]. Chronic Kidney disease reduces the healthy function of kidney and wastes can build to high levels in blood. It leads various abnormalities such as high blood pressure, anemia (low blood count), weak bones; poor nutritional health and nerve damage and it may result to heart and blood vessel disease also. So it is more important to predict the kidney disease at earlier stage. The kidney can be predicted from the record of clinical information. The clinical information can be maintained either manually or systematically. The systematic maintained data can be simpler and powerful than manually maintained data. The clinical data are stored in the database which contains patient information such as age, blood pressure etc., which are used to predict the disease. The features are extracted from the database based on the application. After the future selection the values in each feature are more important to analysis a patient condition. There may be some missing values in database. They are introduced due to manual data entry, software or hardware problem and non response i.e, no information is provided about that data. Even a small percent of missing values in the database leads a bigger disadvantage for further process. There are many researchers introduce many techniques to overcome such problem. The development of Predictive models determines the predictive analytics involving both clinical and nonclinical data which can able to aid in categorize patients with a high probability for missing Hemodialysis treatments [2]. Data mining techniques were applied to analyze the dialysis patients through biochemical data so as to develop a decision support system. In this aspect, the temporal patterns are helpful for doctors to predict hospitalization of hemodialysis patients and to provide proper treatments to avoid hospitalization [3]. Both time series analysis and survival analysis were collectively enables a wide range of risk models in order to evaluate the concordance through discriminatory statistic [4].

In this paper attribute selection technique is used to increase the efficiency of the kidney disease detection. Attribute selection is a process of reducing the dimension of a dataset by eliminating the attributes of less importance and also to select the important attribute to be present in the dataset. Several attribute selection techniques were proposed before namely, Best first Search, Information Gain, Gain Ratio, Best first Search. In [5] Classification models have been framed using classification algorithms along with Wrappersubset attribute evaluator and best first search method even predict and classify non CKD and CKD patients. Each algorithm has significant backlogs namely, requires more time, do not suit for imbalanced dataset and large number of distinct values etc., In order to overcome these problems and also to increase the accuracy of attribute selection, a Priority assigning algorithm is proposed. This algorithm utilizes the notion of "priority", assigning higher priorities to more important features of the data set. It therefore ensures these features or attributes to be present in the reduced or processed data set to

make the classification process efficient. Initially the different types of attributes in the dataset is identified and classified based on priority. The addition of weight values to the attributes are performed automatically with minimal inputs from the user. In order to find the optimal weight value of the attribute, TOPSIS algorithm is utilized. This work selects the attributes based on the priority so that accurate classification can be performed. Algorithm such as Support Vector Machine, Artificial Neural Network, Partial Least Squares Discriminant Analysis (PLS-DA), Weighted Average Ensemble Learning Imputation (WAELI) and Weighted Average Ensemble Learning Imputation with priority assigning algorithm (WAELI-PA).

METHODOLOGY

Weighted Average Ensemble Learning Imputation (WAELI)

The WAELI is used to predict the missing values which include various models are EM, RF, CART and C4.5. These models predict the missing values independently and finally calculate the weight factor of individual model to get a final predictor [6].

TOPSIS Method

TOPSIS is the method used construct best solution and worst solution from the multiple features in the dataset and then set optimal solution from the set of solutions which is closest to the best solution and farthest to the worst solution in TOPSIS method. This is used to estimate important features in the dataset to predict the kidney disease [7].

Construct the theoretical best solution $S^+$and worst solution S^-. $S^+ = \{S_1^+, S_2^+, \dots., S_K^+\}$ and $S^- = \{S_1^-, S_2^-, \dots., S_K^-\}$

$$S_1^+ = \max\{S_{i,j} | i = 1,2, \dots n\} = w_1 \max\{r_{i,j} | i = 1,2, \dots n\} = w_j x_j^+$$

$$S_1^- = \min\{S_{i,j} | i = 1,2, \dots n\} = w_1 \min\{r_{i,j} | i = 1,2, \dots n\} = w_j x_j^-$$

Then for each column in the dataset compute the distance between the features to its theoretical best solution d_i^+ and its theoretical worst solution be d_i^-. It can be calculated by using following equation.

$d_i^+ = \sum_{j=1}^{k} (S_{i,j} - S_1^+)^2 = \sum_{j=1}^{k} {w_1}^2 (r_{i,j} - r_j^+)^2$ $i = 1,2, \dots n$, k is the number of features in the dataset.

$d_i^- = \sum_{j=1}^{k} (S_{i,j} - S_1^-)^2 = \sum_{j=1}^{k} {w_1}^2 (r_{i,j} - r_j^-)^2$ $i = 1,2, \dots n$

Finally the variable of adjacency A_i is computed by:

$$A_i = \frac{d_i^-}{d_i^+ + d_i^-}$$

Weight Self-Production Mechanism

In this mechanism the weight of each features in the dataset is calculated automatically and reasonable weights for different features are calculated. The variable of adjacency A_i is inversely proportional to d_i^+. That means the smaller value of d_i^+, the better the solution. Therefore the objective solution is as follows [8] :

Min $\sum_{i=1}^{n} d_i^+ = \sum_{i=1}^{n} \sum_{j=1}^{k} (r_{i,j} - r_j^+)^2 w_j^2$

$$\sum_{j=1}^{k} w_j = 1$$

Applying the Lagrange multiplier method to the above objective function

$$f_{lagrange}(w, \lambda) = \sum_{i=1}^{n} \sum_{j=1}^{k} (r_{i,j} - r_j^+)^2 w_j^2 - \lambda \sum_{j=1}^{k} w_j - 1$$

According to the theory of the Lagrange multiplier method, the extreme value points must contain in the solutions of the partial derivative equations which is shown below:

$\frac{\delta f_{lagrange}}{\delta w} = 0, \frac{\delta f_{lagrange}}{\delta \lambda} = 0$, so we get:

$$\begin{cases} 2 \sum_{i=1}^{n} (r_{i,j} - r_j^+)^2 w_j - \lambda = 0 \\ \sum_{j=1}^{k} w_j - 1 = 0 \end{cases} \quad j = 1,2,3, \dots . k$$

By solving the above multivariate linear equations, we can finally arrive at:

$$w_j = \frac{1}{\sum_{j=1}^{k} \frac{1}{\sum_{i=1}^{n} (r_{i,j} - r_j^+)^2} \times \sum_{i=1}^{n} (r_{i,j} - r_j^+)^2}$$

Obviously, $w_j \geq 0$. And the weights computed by w_j equation will achieve the minimum value of objective function. Thus, we use w_j equation to determine the weights of different features, and applying the results into the attribute weight vector W of the TOPSIS method.

Proposed Priority Assigning Algorithm

In order to feed all features to a classifier we assign priority for each feature in the dataset then higher priority features are extracted and it will feed into the classifier to predict the kidney disease effectively. It will reduce time consumption during classification. This can be achieved by introducing a new algorithm called priority assigning algorithm. This considers the process of attribute selection to increase the efficiency of the kidney disease detection. Attribute selection is a process of reducing the dimension of a dataset by eliminating the attributes of less importance and also to select

the important attribute to be present in the dataset. The addition of weight values to the attributes are performed automatically with minimal inputs from the user. In order to find the optimal weight value of the attribute, TOPSIS algorithm is utilized. This work selects the attributes based on the priority so that accurate classification can be performed.

We define a method to make the different types of attribute parameters comparable with each other and normalize them. In priority assigning algorithm considered various parameters are cost type parameter, benefit type parameter, Constraint Type Parameter to assign priority and to differentiate the features. In the kidney disease data set cost type parameter are the smaller values but the better they are. They can be normalized by using given equation:

$$r_{i,j} = \frac{x_{i,j}^{max} - x_{i,j}}{x_{i,j}^{max} - x_{i,j}^{min}} \quad \text{i=1,2….n}$$

Where,

$x_{i,j}^{max}$ = Maximum value of the corresponding attributes parameters among all the columns in the dataset

$x_{i,j}^{min}$ = Minimum value of the corresponding attributes parameters among all the columns in the dataset

$x_{i,j}$= is the attribute value of the current column.

In the kidney disease dataset the benefit type parameter the larger the better. It can be normalized by the given below equation:

$$r_{i,j} = \frac{x_{i,j} - x_{i,j}^{min}}{x_{i,j}^{max} - x_{i,j}^{min}} \text{ i=1,2….n}$$

In the kidney disease dataset the constraint type parameter must satisfy the specific threshold. It can be normalized by following equation:

$$r_{i,j} = \begin{cases} 1 & x_{i,j} \leq t \\ 0 & x_{i,j} \geq t \end{cases}$$

Where, t= specific threshold.

The priority algorithm maintains queue of features and the features are ordered in descending order of priority. Normal queue is FIFO so high priority features are in head of the queue and it will be given to the classifiers to classify the data and to predict the kidney disease effectively. In this algorithm, the decision process is dived into two steps: the features which cannot meet the constraints of objective function are removed from consideration then all the remaining features are selected to form feasible solution set for the dataset. In the second step, an autonomous weight production mechanism is designed to determine the weights of the multiple attributes, and utilize the TOPSIS method to find the optimal feature in the set of multiple features in the dataset. If the feasible solution for the features is empty, then this algorithm will initiate the adjustment process will remove all low priority features from the dataset and the dataset is recalculated again and select the optimal solution which contain high priority features is found. The optimal solution gives the high priority (important) features in the dataset.

```
Priority Assigning Algorithm
Initialize ( ) //initialize the algorithm
{
determine the priorities of different features in the dataset;
queue the features in descending order of priorities;
}
Procedure Distribution ( )
//distribute the feature to the appropriate access dataset
{
while (the queue is not empty)
{
get and delete the first feature from the queue;
calculate the feasible solution set for the feature according to the Constraints;
if (the feasible solution set is not empty){
//normalize the attribute parameters
Normalization ( );
//produce the weights of the different attribute
Weight_SelfProduction ( );
//find the optimal solution for the feature
TOPSIS ( );
}
else
Adjustment ( );
{
recalculate the feasible solution set for the feature according to the Constraints;
}
}
}
```

Figure 1: Proposed Priority Assigning Algorithm

EXPERIMENTAL ANALYSIS

Social Accuracy is defined as the proportion of true positives and true negatives among the total number of results obtained. Here we introduce a new algorithm called priority assigning algorithm which assigns high priority for more important features. Which makes the classification process efficient and kidney disease can be predicted efficiently.

The values of accuracy, precision, recall and F-measure values of WAELI and WAELI-FPA are tabulated is shown below using Matlab 2013. Accuracy is evaluated as,

$$Accuracy = \frac{(True\ positive + True\ negative)}{(True\ positive + True\ negative + False\ positive + False\ negative)}$$

The accuracy of proposed technique can be evaluated based on how attribute are selected based on priority and it is given to various classifiers to predict the kidney disease. In Figure 2 illustrates the accuracy between WAELI and WAELI-FPA. Weighted Average Ensemble Learning Imputation technique fills the missing values in the database. The missing values are imputed either by using single value imputation or multiple value imputation. In single value imputation the missing values are predicted values and it is replaced with single value by using Expectation-Maximization and Random Forest. In multiple value imputation the missing values are imputed replacing missing values with more than one random value by using Random Forest, CART and C4.5.

The values have been calculated using existing algorithm such as SVM, ANN, PLSDA and MFPT respectively. The missing values in the dataset are predicted using WAELI and the entire features in the dataset is given to different classifiers like SVM, ANN, PLS-DA, MFPT and it predict the patient with kidney disease. Instead of giving all attributes to classifiers we used priority algorithm to select the high priority features then those features are given to the classifiers it predict the kidney disease effectively.

The proposed priority assigning algorithm uses TOPSIS method to classify the best and worst solution and weight self production method which determines the weights of the different attributes, that greatly influence the decision making process. In figure 2 for WAELI-SVM the accuracy rate is found to be 73%, WAELI-FPA-SVM shows 78.5%, WAELI-ANN shows 73%,WAELI-FPA-ANN 78%,WAELI-PLSDA shows 73%,WAELl –FPA-PLSDA shows 78.8%,WAELI-MFPT shows 73% and WAELI-FPA-MFPT shows 78.8% approximately. This shows WAELI-FPA gives high accuracy than WAELI as shown below in figure.

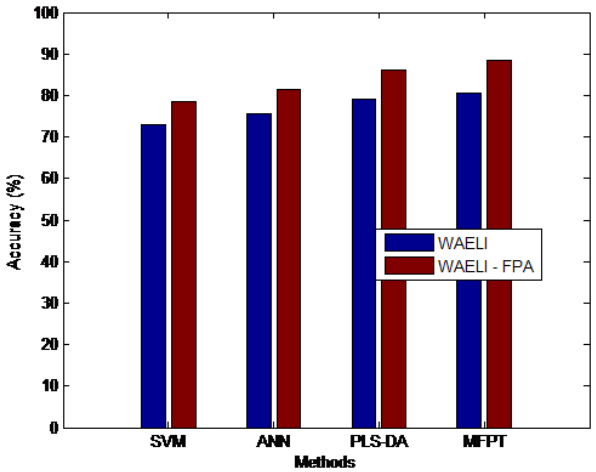


Figure 2: Accuracy value between WAELI and WAELI-FPA

Precision value is evaluated according to the relevant information at true positive prediction, false positive.

$$Precision = \frac{True\ positive}{(True\ positive + False\ positive)}$$

The precision of proposed technique can be evaluated based on how attribute are selected based on priority and it is given to various classifiers to predict the kidney disease. Figure 3 illustrates precision between WAELI and WAELI-FPA. For WAELI-SVM the precision rate is found to be 75.5%, WAELI-FPA-SVM shows 81.5%, WAELI-ANN shows 75.58%,WAELI-FPA-ANN 81.42%,WAELI-PLSDA shows 75.70%,WAELl –FPA-PLSDA shows 81.37%,WAELI-MFPT shows 75.64% and WAELI-FPA-MFPT shows 81.40% approximately. As a result the precision rate of proposed WAELI_FPA shows better performance than compared with WAELI algorithm respectively.

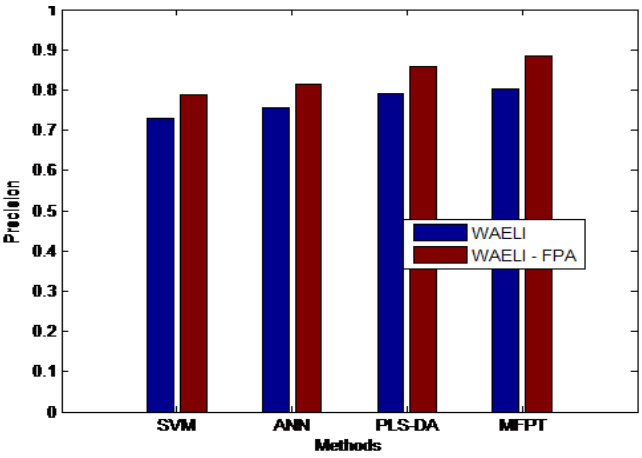


Figure 3.: Precision value between WAELI and WAELI-FPA

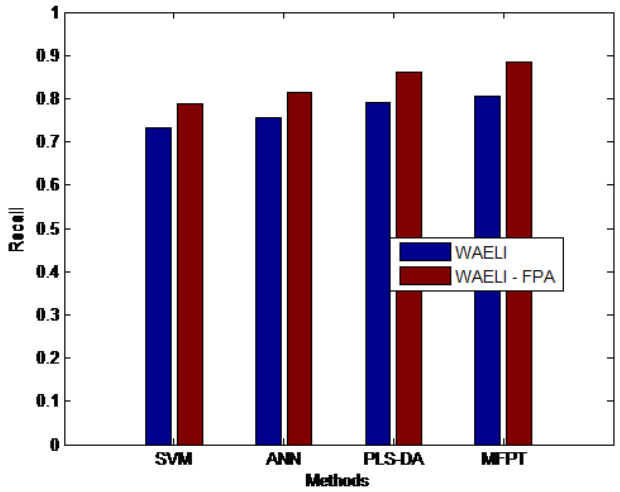


Figure 4: Recall Value between WAELI and WAELI-FPA

The Recall value is evaluated according to the retrieval of information at true positive prediction, false negative.

$$Recall = \frac{True\ positive}{(True\ positive + False\ negative)}$$

Figure 3 illustrates precision between WAELI and WAELI-FPA. For WAELI-SVM the recall value of precision rate is found to be 79%, WAELI-FPA-SVM shows 86%, WAELI-ANN shows 79%,WAELI-FPA-ANN 85%,WAELI-PLSDA shows 79%,WAELl –FPA-PLSDA shows 86%,WAELI-MFPT shows 79% and WAELI-FPA-MFPT shows 85% approximately.

F-measure is calculated from the precision and recall value. It is calculated as:

$$f - measure = 2 \times \left(\frac{precision \times recall}{precision + recall}\right)$$

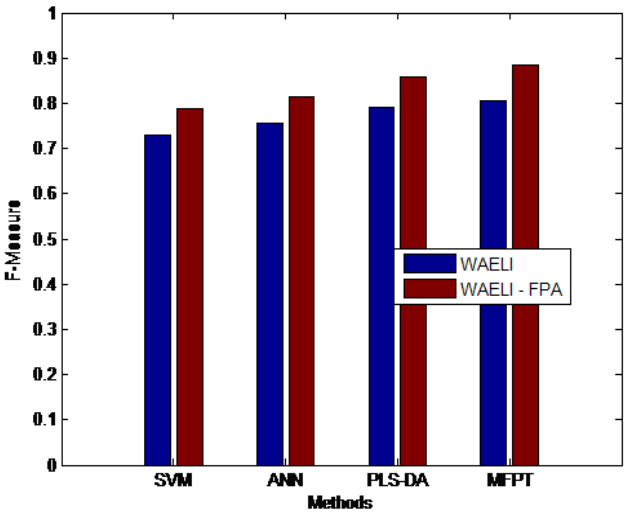
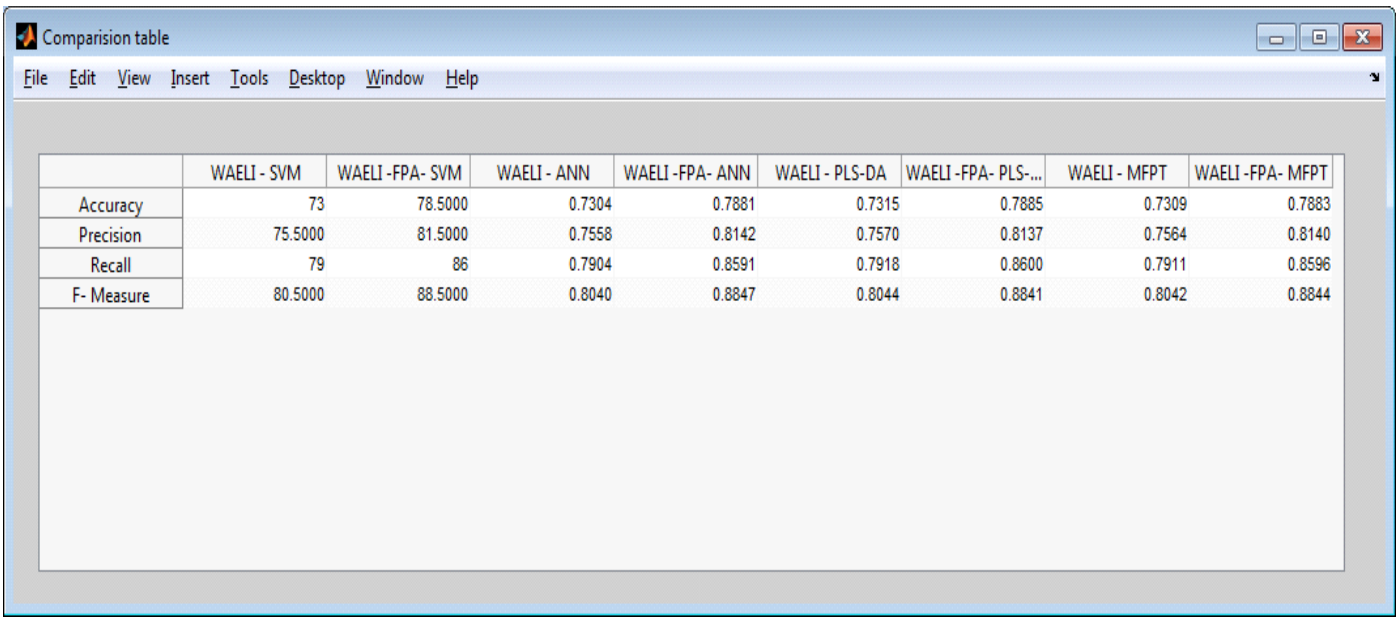


Figure 5: f-measure between WAELI and WAELI-FPA

The f-measure of proposed technique can be evaluated based on how attribute are selected based on priority and it is given to various classifiers to predict the kidney disease. Figure 3 illustrates precision between WAELI and WAELI-FPA. For WAELI-SVM the F-measure value is found to be 80.4%, WAELI-FPA-SVM shows 88.5%, WAELI-ANN shows 80.4%,WAELI-FPA-ANN 88.4%,WAELI-PLSDA shows 80.4%,WAELl –FPA-PLSDA shows 88.4%,WAELI-MFPT shows 80.4% and WAELI-FPA-MFPT shows 88.4% approximately.



	WAELI - SVM	WAELI -FPA- SVM	WAELI - ANN	WAELI -FPA- ANN	WAELI - PLS-DA	WAELI -FPA- PLS-...	WAELI - MFPT	WAELI -FPA- MFPT
Accuracy	73	78.5000	0.7304	0.7881	0.7315	0.7885	0.7309	0.7883
Precision	75.5000	81.5000	0.7558	0.8142	0.7570	0.8137	0.7564	0.8140
Recall	79	86	0.7904	0.8591	0.7918	0.8600	0.7911	0.8596
F- Measure	80.5000	88.5000	0.8040	0.8847	0.8044	0.8841	0.8042	0.8844

Figure 6: Performance Analysis of Proposed WAELI and WAELI_FPA Algorithm

While using existing algorithm like SVM, ANN, PLS-DA and MFPT algorithms the performance is seems to be low in predicting the kidney disease. So the classifiers are combined with Weighted Average Ensemble Learning Imputation (WAELI) technique. On account of this in order to predict much better priority assigning algorithm has been proposed. Based on this, the classifiers are combined with WAELI-FPA (Feature based Priority Assigning Algorithm). Obviously WAELI-FPA technique gives better results compared to the previous observations.

CONCLUSION

Then introducing priority assigning algorithm to assign priority for each features in the dataset then higher priority features are carried over for classification process. This makes classification process more efficient and time consumption for classification will be reduced. From the experimental results, it is proved that the proposed WAELI and WAELI FPA effectively predict the kidney disease with high accuracy, precision, recall, F-measure and with low RMSE and MAE values compared with the existing system.

REFERENCES

[1] About Chronic Kidney Disease - The National Kidney Foundation, https://www.kidney.org/atoz/content/about-chronic-kidney-disease, 2017.

[2] Yue Jiao, PhD, Dan Geary, MHA, Sheelal , Peter Kotanko, Development and Testing of Prediction Models for End Stage Kidney Disease Patient Nonadherence to Renal Replacement Treatment Regimens Utilizing Big Data and Healthcare Informatics, 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) Pages 1721-1721.

[3] Jinn-Yi Yeh, Tai-Hsi Wu, Chuan-Wei Tsao, Using data mining techniques to predict hospitalization of hemodialysis patients Decision Support Systems 50 (2011) 439–448.

[4] Adler Perotte, Rajesh Ranganath, Jamie S Hirsch, David Blei, Noemie Elhadad Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis, Research and Applications, 872-880.

[5] Naganna Chetty Kunwar Singh Vaisla, Sithu D Sudarsan Role of Attributes Selection in Classification of Chronic Kidney Disease Patients Computing,International Conference on Communication and Security (ICCCS), 2015.

[6] S.Dilliarasu and R. Thirumalaiselvi, A novel imputation method for effective prediction of coronary Kidney disease, Second IEEE International Conference on Computing and Communication Technologies,2017, 127-136

[7] Bondor CI, Kacso IM, Lenghel AR, Mureşan A, Hierarchy of risk factors for chronic kidney disease in patients with type 2 diabetes mellitus. Conf Procedings of IEEE International Computer Communication, 103-106.

[8] Yi Sun, Yuming Ge, Jue Yuan, Jihua Zhou Stephen, Herborn Dongdong Chen, PAWES: A Flow Distribution Algorithm Based on Priority and Weight Self-Production IEEE Communications Society subject matter experts for publication in the WCNC 2009 proceedings.1-6.