Optimal design of quadratic support vector machine based end stage renal disease predictor model

S Sathyadeepa¹, Dr. T. Sivakumar¹

¹RVS College of Arts & Science, Department of Electronics, Coimbatore, Tamilnadu- 641 402, India.

E-mail: sat22mahe@gmail.com

Abstract: At present kidney failure has become one such important paradigm for higher mortality rate. Chronic kidney disease (CKD) is a severe kidney infection, and it has different stage of infection with final condition is being called as End Stage Renal Disease (ESRD). At this juncture patient are supported with dialysis or kidney transplant. The process requires time to analyze, diagnose the different parameters in determining CKD. The study involves using machine learning techniques for predicting the stages involved in infection and prevent the disease progression. There are four different stages of kidney disease, patient above 12 years of age are analyzed with 24 different input parameters. Around 160 subject’s parameters were taken into the database. Important feature includes serum creatinine, blood pressure, bacterial infection, pedal edema, urine sugar and sodium level in the blood are included. Machine learning algorithm was developed using MATLAB 2018 to predict the early end stage of CKD. Algorithms like Fine Decision Tree (FDT), Quadratic Support Vector Machine (Quad-SVM) and Linear Discriminant Analysis (LDA) are the three main algorithms used for designing the predictor model. Classification Learner Toolkit helps is determining the accurate model and Curve Fitting Toolbox helps in curve fitting over the given data set. After analyzing the system with different predictor model, the model with high accuracy and less training time is chosen. The proposed work also incorporates an e-message alert system linking the nephrologists in case of any emergency wirelessly. It enables high speed prediction and diagnosis process for treatment of subject under high risk factor. The threshold value is continuously monitored and in case if the value exceeds the safer limit, a notification is sent to the physician in proper time.

Keywords— Chronic Kidney Disease, Machine learning algorithms, Fine Decision Tree, quadratic Support Vector Machine, Linear Discriminant Analysis, Classification Learner, Curve Fitting

I. INTRODUCTION

In the present scenario disease prediction and evaluation of renal failure has become so common. With regular lab tests it is possible to predict the kidney disease and there is a normal treatment which can reduce the inflammation and maintain the glomerular filtration rate (GFR) rate at the prescribed limit. GFR indicates the functioning capacity of CKD. Serum creatinine and urine is used for kidney disease diagnosis. Different methods are used for screening out of which ultrasound screening is the most common method to diagnose any abnormality [1]. Basically, the treatment slows down the progression of kidney disease, it becomes essential to design a better predictor model for identifying patient with high risk of kidney failure and avoid such condition with early diagnosis [2].

An equation was developed to predict the risk factor with four different variables. Ratio of Albumin level and creatinine level and GFR is used in the equation to understand the progression of the disease and also predicting the end stage. The prediction takes place before 2-5 years of kidney transplant and predictors are determined [3]. During the end stage of the
infection, the functioning of the kidney is reduced by 15% when compared to the normal functioning kidney. In this stage the filtration capacity is drastically reduced [4].

A model is designed, and the success rate is achieved around 80%, the construction of the model involves decision tree algorithm technique classification for predicting the success rate of the diagnosis. 20% of the data is used for examination and remaining 80% of the data is used for model construction. Different subset involving the testing data set are classified and this process is repeated till the new subsets are formed. The final testing data set is applied for remaining unused data [5-8]. Drugs especially used for hypertension, haemoglobin level, drugs treating albumin deficiency, blood pressure and age is mainly used as a parameter for prediction model. Also, there is an additional set of parameters like smoking (male subject), longer period of diabetes (female subject), insulin level (young male), GFR (young female) and higher body mass index are an important parameter for better prediction model [9].

In the biomedical sector, machine learning techniques are booming up for prediction model design. Different diseases are being predicted by these techniques which helps the human society to predict the onset of the disease and step forward for preventing it. The work done in [10] clearly depicts how big data comes to aid the machine learning algorithms for disease prediction. Even kidney diseases are predicted using various parameter and detailed investigation report was submitted [12]. Feed-forward neural networks were used to classify the stages involved in CKD along with the help of Logistic Regression and deep learning. Also kernel function-based prediction model was developed for this disease prediction [13]. Naïve Bayes algorithm along with Decision tree provided better results. On contrary to that even SVM and K-Nearest Neighbor were modelled to predict the disease [14].

Back propagation neural network was used to improve the model prediction characteristics. Random forest is a technique used in various application of disease prediction. It was concluded that better than neural network-based prediction model deep learning-based classifier provided higher accuracy of around 86% [15-17]. Adaptive boosting algorithm unlike Random subspaces for diagnosing the onset of CKD provided better response, ensemble classifier provided better resolution among other learning classifiers for the output class [18]. A comparative study was made between different classifiers and machine learning techniques and the best performance was elucidated from SVM and decision tree [19]. Also, J48 algorithm was compared with Naïve Bayes algorithm and it was observed that, prediction time better for J48. Along with it even minimal sequential optimization technique was used against adaptive boosting algorithm for improving the accuracy [20]. In the proposed work, very high level of accuracy was obtained using Quad-SVM technique which can be compatible for different conditions.

II. METHODOLOGY

A predictor model for determining if a subject is highly susceptible to renal failure. Here three model prediction technique like Fine Decision Tree (FDT), Quadratic Support Vector Machine (Quad-SVM) and Linear Discriminant Analysis (LDA).

A. Fine Decision Tree

Classification kind of problems and regression-based problems can be easily solved with this technique. The tree like structure is used for solving any kind of problem, each leaf node
represents classes and further the internal node corresponds to the parameters. The complete training set used in the model prediction is assumed to be the root of the tree. There are 24 features, but each feature must be categorical it cannot be character type. Model construction includes discretizing the data prior to prediction and with the help of statistical method internal node or root is corresponded as features. Decision tree flow chart is shown in Fig 1, haemoglobin value is taken as the parent node and creatinine level and specific gravity is taken as child nodes, from which further bifurcation occurs.

![Decision Tree for CKD prediction](image)

**Fig 1.** Decision Tree for CKD prediction

### B. Quadratic Support Vector Machine (Quad-SVM)

SVM was the tool to develop an efficient algorithm for non-linear classification problems. In earlier methods local minima and local maxima problem was yet to be solved. Quadratic optimization comes to rescue for such problems where the kernel function used for SVM relies on quadratic paraboloid function. If the function is linearly separable then hyper plane margin is optimized whereas if the output cannot be linearly separable the original data is mapped over the new space defined by the kernel function. To achieve maximum accuracy of prediction quadratic kernel function is used which is shown in the Fig 2. ‘0’ is represented as green star and ‘1’ is represented as red circle. D+ and D- is the margin of the hyperplane. As the boundary of the margin increases the data points are included. The curve of the hyperplane is not a straight line instead it is a parabola representing quadratic equation.
C. Linear Discriminant Analysis (LDA)

The assumption made on Gaussian distribution as allowing the classes from different category to generate data is the first step for model prediction. Unlike the kernel function here fitting function is used for estimation of parameters used in each class. Each output class is divided as person having CKD and person not having CKD. Basically, the output class represents two states such as ckd or notckd which means that whether the person is having high risk towards CKD or not. This method has two different classifiers namely linear and quadratic discriminant analysis.

Suppose that \( Y \) is an \( axb \) matrix, where ‘\( b \)’ is the total count of predictor variable and ‘\( a \)’ is the total count of data points. Also, \( Y_i \) and \( Y_j \) are \( i^{th} \) row and \( j^{th} \) column of the \( Y \) matrix respectively. The number of observations in \( m^{th} \) group is represented by \( C_m \), that is a subset of the following equations (1)-(4).

\[
a_m = |C_m| \quad (1)
\]
\[
a = \sum_{m=1}^{n} a_m \quad (2)
\]

An estimate for the output target within-class and covariance matrix \( \Sigma_w \) and between classes \( \Sigma_t \) is represented as the following,

\[
\Sigma_w = \frac{1}{a} \sum_{m=1}^{m} \Sigma(Y_i - \mu_m)(Y_i - \mu_m)^T \quad (3)
\]
\[
\Sigma_t = -\frac{1}{a} (Y^TY) - \Sigma_w = \frac{1}{a} \sum_{m=1}^{m} \mu_m \mu_m^T \quad (4)
\]

Actually, \( \mu_m \) is the mean sample of the class ‘\( m \)’. Finally, a vector \( \beta_m \) of class ‘\( m \)’ needs to be maximized which is given in the equation (5).

\[
\max(\beta_m(\beta_m^T \sum \beta_m - B_m(\beta_m))) \quad (5)
\]
The condition is such that $\beta^T_m \sum \omega \beta_m \leq 1$, where $B_m$ is the penalty function used for the $m^{th}$ discriminant vector. Here in the condition positive definite value of $\sum \omega$ is used in the equation (5) for maximization.

### Table 1. Illustration of parameters for CKD prediction model.

| S. No. | Features                  | Description                              | Type of data |
|--------|---------------------------|------------------------------------------|--------------|
| 1      | Age                       | Subject’s Age (12-79 years)              | Integer      |
| 2      | RedBloodCells             | Red blood cells count whether normal or abnormal (normal-0 and abnormal-1) | Binary       |
| 3      | SpecificGravity           | Specific gravity value of urine (1.007-1.010) | Integer      |
| 4      | BloodGlucoseRandom        | Random blood glucose level               | Integer      |
| 5      | Potassium                 | Hyperkalaemia level                      | Integer      |
| 6      | PusCellClumps             | Pus cell clumps are present (present-1, not present-0) | Binary       |
| 7      | WhiteBloodCellCount       | White blood cell count for disease progression | Integer     |
| 8      | SerumCreatinine           | Creatinine level in urine                | Integer      |
| 9      | PedalEdema                | Swelling around legs (yes-1, no-0)       | Binary       |
| 10     | DiabetesMellitus          | Type 2 diabetes (yes-1, no-0)            | Binary       |
| 11     | BloodPressure             | Blood pressure level                     | Integer      |
| 12     | CoronaryArteryDisease     | Presence of heart disease (yes-1, no-0)  | Binary       |
| 13     | Anaemia                   | Presence of anaemia (yes-1, no-0)        | Binary       |
| 14     | BloodUrea                 | Blood urea level                         | Integer      |
| 15     | RedBloodCellCount         | Red blood cell count (abnormal-1, normal-0) | Binary       |
| 16     | PusCell                   | Formation of pus cell in kidney (abnormal-1, normal-0) | Binary       |
| 17     | PackedCellVolume          | Packed cell volume value                 | Integer      |
| 18     | Hypertension              | Presence of hypertension (yes-1, no-0)   | Binary       |
| 19     | Appetite                  | Appetite condition (poor-1, good-0)     | Binary       |
| 20     | Haemoglobin               | Haemoglobin level                        | Integer      |
| 21     | Albumin                   | Albumin Level                            | Integer      |
| 22     | Sugar                     | Sugar level in urine (Five conditions)   | Integer      |
| 23     | Bacteria                  | Bacterial infection (present-1, not present-0) | Binary       |
| 24     | Sodium                    | Renal sodium content                     | Integer      |

There is total 24 features for model prediction and 1 output target which has two condition either 1- person having CKD and 0-person not having CKD as shown in Table 1. parameters like age, red blood cell count, specific gravity and serum creatinine level is a predominant parameter among all. It can be better understood with the help of correlation matrix. By designing a predictor model, it can be used to easily identify the subjects with CKD and carry out further tests for confirmation and treatment.

### III. RESULT AND DISCUSSION

It involves various methods used in determining the final model and evaluating the best model among them. The steps involved in the CKD predictor model are discussed below:
A. Data Pre-Processing

The data that is received from the health centres for model design is highly inefficient in nature as it may contain missing data, infinite values data and data is character format. In order to make the data more viable for its use, it becomes eminent to process the data. In MATLAB data pre-processing involves replacement of data with unimportable values to Nan (infinite) and blank data as Nan. In this way the data that are dummy can be transformed to something quite useful. In MATLAB there is yet another feature that enables the user to change the ‘char’ type of data into binary data which involves abnormal-1 and normal-0, yes-1 and no-0, present-1 and notpresent-0 and poor-1 and good-0.

B. Data balancing

In this stage, the balanced data sets for model prediction design. The total number of available data set for the CKD prediction is around 160. Out of which 110 data was classified under the output class ‘0’ and 50 data was classified under the output class ‘1’ which is clearly shown in Fig 3. In MATLAB there is a feature known as random permutation which enables in choosing the random data from the complete database to balance the complementary data. After balancing the number of data sets belonging to the output class ‘0’ and ‘1’ are the same which is 43. Hence, the total data set used for model prediction is 86 as shown in Fig 4. This technique is essential as the data if unbalanced can give biased output which will be similar to the high number output class. It makes the model prediction more accurate.

![Fig 3: Histogram of output class of unbalanced data](image1)

![Fig 4. Histogram of output class of balanced data](image2)
C. Principal Component Analysis (PCA)

In MATLAB, a technique known as PCA for removing redundant features is used. There are total of 24 parameters, which interacts with each other as well as with the output class. It must be noted that the interaction between the input and output feature must be highly whereas between input and input feature it must be uncorrelated. The correlation factor lies in between the range of (0-1).

If the value is less than 0.5, the correlation is low and recommended. If the input-input correlation factor is high and the input-output correlation factor is low then such kind of parameters can be removed. This process is done by the inbuilt function available in the software. Fig 5. clearly explains the correlation among the inputs and value of correlation factor for all the 23 parameters are also given. For example, the albumin level and packed cell volume is highly and negatively correlated. Basically, negative correlation means that if albumin level rises then level of packed cell volume decreases in a subject. Magnitude of negative correlation is always considered. And the correlation factor is 0.79 as shown in the Fig 5, it clearly indicates that input-input correlation is high hence it can be preferred to be removed.

![Fig 5. Input-Input Correlation Matrix for CKD](image)

D. Decision Tree

In this method classification technique to group the data sets. Actually, the tree is limited certain number of divisions, the maximum split that can occur is 100 and the criteria used for splitting is known as Gini’s diversity index. All the 24 parameters are used for model prediction by disabling the PCA option. Scatter plot sown in Fig 6. explains how the two important features like age and serum creatinine vary. In the Fig 6. red dot represents the ‘yes’ class and blue dot represents the ‘no’ class. Blue cross represents the wrongly classified ‘no’ class member under ‘yes’ class.
Confusion matrix is shown in Fig 7, X-axis of the confusion represents predicted class and Y-axis of the confusion matrix represents true class. From Fig 7, it is clear that all the 43 data set belonging to the ‘no’ class is rightly grouped. Then, 40 data set belonging to the ‘yes’ class is correctly classified and 3 data set is wrongly classified as ‘no’ in the predicted model.

Receiver Operating Characteristic also known as ROC curve depicts the characteristics of the model. ROC is determined by the Area Under Curve (AUC), the minimum value of AUC can be 0 and maximum can be 1. In case of Decision tree it is around 0.97 closer to 1 as shown in Fig 8(a) and 8(b) for ‘no’ and ‘yes’ class respectively, that means higher accuracy and better prediction.
E. Linear Discriminant Analysis

This method uses vector method for classification and full covariance structure was used. During quadratic discriminant analysis one of the predictors remained constant without changing. Hence such predictors must be removed, and the covariance structure used must be changed for using quadratic discriminant analysis. Scatter plot is shown in Fig 9, as it depicts the parametric variation among the age and serum creatinine for prediction of CKD in the subject. Here also red dot represents the ‘yes’ class and blue dot represents the ‘no’ class. The blue cross represents the wrongly classified data set from the ‘no’ class under ‘yes’ class. It can
be observed that serum creatinine level is high in among the age group 50-70 depicting high risk of CKD.

Confusion Matrix for LDA is shown in Fig 10. It can be observed that 43 data set belonging to ‘no’ class is rightly classified under ‘no’ class. Then 41 data set belonging to the ‘yes’ class is rightly classified under ‘yes’ class but 2 data set belonging to ‘yes’ class is wrongly classified under ‘no’ class.

ROC characteristics is shown in Fig 11(a) and 11(b) for ‘no’ class and ‘yes’ class respectively. The current classifier for both the classes are different. The AUC value is 1 approximately. It means very high accuracy value and better prediction model.
Fig 11(a). Discriminant Analysis Model ‘no’ Class- ROC Curve

Fig 11(b). Discriminant Analysis Model ‘yes’ Class- ROC Curve

F. Quad-SVM

This method can be used for classification as well as regression. Here output class must be classified in ‘yes’ and ‘no’ class. The ‘yes’ class indicates person having CKD and ‘no’ class indicates person not having CKD. Quad-SVM uses quadratic kernel function for non-linear mapping of output class. Scaling of kernel function is done automatically, and one constraint is applied. One vs one multiclass method is used for classification of any data set into binary classification such as ‘yes’ and ‘no’ class. Scatter plot is shown in Fig 12, it can be observed that how the serum creatinine level varies with age of the person. Here also the red and blue dot
represents ‘yes’ and ‘no’ class respectively. Blue cross represents data set from ‘yes’ class is wrongly classified under ‘no’ class.

Fig 12. Quad-SVM Model-Scatter Plot

Confusion matrix for Quad-SVM shown in Fig 13 depicts that it around 43 data set belonging to the ‘no’ class is rightly classified under ‘no’ class and 42 data sets from ‘yes’ class is rightly classified under ‘yes’ class whereas 1 data set is wrongly classified under ‘no’ class.

Fig 13. Quad-SVM Model- Confusion Matrix

ROC characteristics for Quad-SVM model is shown in Fig 14(a) and 14(b) for ‘no’ class and ‘yes’ class respectively. The value of AUC for both the output class is 1 approximately.
which means that the accuracy level is very high. The current classifier values of two classes differ by minimum value.

![ROC Curve](image)

**Fig 14(a).** Quad-SVM Model ‘no’ class- ROC Curve

**Fig 14(b).** Quad-SVM Model ‘yes’ class- ROC Curve

Finally, the comparison among different method is done and tabulated in the Table 2. It can be concluded that Quad-SVM technique has highest accuracy of 98.8% among all other techniques. It has better training time of 0.8 sec approximately. And AUC is 1, it can be said that Quad-SVM is the most preferable computational algorithm for designing the predictor model with very minimal false classification. Then Linear DA can be considered better for low
prediction time which is around 860 obs/sec. even its accuracy is higher than Fine tree but lower than Quad-SVM.

| S. No. | Accuracy (%) | Prediction Speed (obs/sec) | Training Time (s) | Area Under Curve | Features |
|--------|--------------|-----------------------------|-------------------|-----------------|----------|
| Fine Tree | 96.5          | 550                         | 4.014             | 0.97            | 24       |
| LDA    | 97.7          | 860                         | 2.5716            | 1               | 24       |
| Quad-SVM | 98.8         | 2500                        | 0.8221            | 1               | 24       |

**IV. CONCLUSION**

Prediction of chronic ailments like renal failure has become quite popular in the domain of Artificial Intelligence. It is challenging to design a predictor model with minimal data set and overcoming the problems related to over as well as under sampling. Efficiency of any predictor model is eminent, improving it by computational techniques and analyzing the characteristic curves becomes mandatory. Among the proposed algorithms for design, Quad-SVM provides better training time and higher accuracy compared to other models.

Quality of data and sampling size is also an important feature and a key parameter for better prediction. Higher accuracy is attained with the help of data pre-processing and data balancing without losing the actual property of the data set. The future scope of the research work incorporates parameter like bilirubin for better prediction of the disease with the help of multi-parameter estimation. The proposed work predicts the person with high risk of CKD, hence further medical processes must be carried out in time to avoid disease progression. This kind of disease prediction technique not only saves time but also directs the physicians to take the treatment to the higher level so that many lives could be saved.

**V. REFERENCES**

[1] Zvi Segal, Dan Kalifa, Kira Radinsky, Bar Ehrenberg, Guy Elad, Gal Maor, Maor, Muhammad Tibi and Gideon Koren “Machine learning algorithm for early detection of end-stage renal disease”, BMC Nephrology, Vol. 518, No. 21, pp. 1-10, 2020.

[2] Guidelines: “Chronic kidney disease in adults: assessment and management,” NICE Clinical guidelines [CG182] published on: 23 July 2014.

[3] Major RW, Shepherd D, Medcalf JF, “The kidney failure risk equation for prediction of end stage renal disease in UK primary care: an external validation and clinical impact projection cohort study,” PLoS Med. 2019.

[4] Robert Thomas, Abbas Kanso and John R. Sedor "Chronic Kidney Disease and Its Complications", Primary Care: Clinics in Office Practice Vol. 35, No. 2, pp. 329-344, 2008.

[5] Hu TK. “Random decision forests,” 3rd International Conference on Document Analysis and Recognition; 1995.

[6] Chen T and Guestrin C. “XGBoost: a scalable tree boosting system,” KDD conference, 2016.

[7] Shalev-Scheutz S, Ben-David S, “Understanding machine learning; from theory to algorithms,” Cambridge: Cambridge University Press; 2014.
[8] Zadronzny B, Elkan C, “Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers,” In: Proceedings of the 18th International Conference on Machine Learning. San Francisco: ICML 01; 2001. p. 609–16.

[9] Wan EYF and Fong DYT. “Prediction of new onset of end stage renal disease in Chinese patients with type 2 diabetes mellitus – a population based retrospective cohort study,” BMC Nephrol. Vol. 18, No. 1, pp. 257, 2017.

[10] Vinitha S., Sweetlin S., Vinusha H. and Sajini S. “Disease prediction using machine learning over big data”, Comput. Sci. Int. J. (CSEIJ), Vol. 8, pp. 1-8, 2018.

[11] Subas, A., Alićkovic, E., Kevric, J.: “Diagnosis of chronic kidney disease by using random forest”, CMBEBIH, Vol. 62, pp. 589–594, 2017.

[12] Imran, A.A., Amin, M.N., Johora, F.T. “Classification of chronic kidney disease using logistic regression, feedforward neural network and wide and deep learning,” In: 2018 International Conference on Innovation in Engineering and Technology (ICIET), pp. 1–6, 2018.

[13] Wibawa, H.A., Malik, I., Bahtiar, N.: “Evaluation of kernel-based extreme learning machine performance for prediction of chronic kidney disease”, In: 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS), pp. 1-4, 2018.

[14] Radha, N. and Ramya, S., “Performance analysis of machine learning algorithms for predicting chronic kidney disease,” Int. J. Comput. Sci. Eng. Open Access 3, Vol. 3, No. 8, pp. 72–76, 2015.

[15] Ramya, S., Radha, N. “Diagnosis of chronic kidney disease using machine learning algorithms,” Int. J. Innov. Res. Comput. Commun. Eng. Vol. 4, No. 1, pp. 812–820, 2016.

[16] Charleonnan, A., Fuafang, T., Niyomwong, T., Chokcheypat, W. and Suwanwach, N. “Predictive analytics for chronic kidney disease using machine learning techniques,” In: 2016 Management and Innovation Technology International Conference (MITicon), pp. 80–83, 2016.

[17] Gunarathne, W.H.S.D., Perera, K.D.M., and Kahandawaarachchi, K.A., “Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD),” In: 2017 IEEE 17th (BIBE), pp. 291–296, 2017.

[18] Basar, M.D., Akan, A. “Detection of chronic kidney disease by using ensemble classifiers,” In: 2017 10th International Conference on Electrical and Electronics Engineering (ELECO), pp. 544–547 2017.

[19] Tekale, S., Shingavi, P., Wandhekar, S., Chatorikar, A. “Prediction of chronic kidney disease using machine learning algorithm,” Int. J. Adv. Res. Comput. Commun. Eng. Vol. 7, No. 10, pp. 92–96, 2018.

[20] Sisodia, D.S., Verma, A. “Prediction performance of individual and ensemble learners for chronic kidney disease,” In: 2017 International Conference on Inventive Computing and Informatics (ICICI), pp. 1027–1031, 2017.