Predictive Analytics of Chronic Kidney Disease using Machine Learning Algorithm

S. Pitchumani Angayarkanni

Abstract*: According to the health statistics of India on Chronic Kidney Disease (CKD) a total of 63538 cases has been registered. Average age of men and women prone to kidney disease lies in the range of 48 to 70 years. CKD is more prevalent among male than among female. India ranks 17th position in CKD during 2015[1]. This paper focus on the predictive analytics architecture to analyse CKD dataset using feature engineering and classification algorithm. The proposed model incorporates techniques to validate the feasibility of the data points used for analysis. The main focus of this research work is to analyze the dataset of chronic kidney failure and perform the classification of CKD and Non CKD cases. The feasibility of the proposed dataset is determined through the Learning curve performance. The features which play a vital role in classification are determined using sequential forward selection algorithm. The training dataset with the selected features is fed into various classifier to determine which classifier plays a vital and accurate role in detection of CKD. The proposed dataset is classified using various Classification algorithms like Linear Regression(LR), Linear Discriminant Analysis(LDA), K-Nearest Neighbour(KNN), Classification and Regression Tree(CART), Naive Bayes(NB), Support Vector Machine(SVM), Random Forest(RF), eXtreme Gradient Boosting(XGBoost) and Ada Boost Regressor (ABR). It was found that for the given CKD dataset with 25 attributes of 11 Numeric and 14 Nominal the following classifier like LR, LDA, CART,NB,RF,XGB and ABR provides an accuracy ranging from 98% to 100% . The proposed architecture validates the dataset against the thumb rule when working with less number of data points used for classification and the classifier is validated against under fit, over fit conditions. The performance of the classifier is evaluated using accuracy and F-Score. The proposed architecture indicates that LR, RF and ABR provides a very high accuracy and F-Score.

Index Terms: Chronic Kidney Disease, Classifier, Logistic Regression, Ada Boost Regressor and Performance metrics..

I. INTRODUCTION

Chronic kidney disease is considered as a serious problem because if not predicted and treated on time it leads kidney failure and hence increases the mortality rate. It is considered as one of the important health issue which needs early detection for successful treatment. Many factors like diabetics, blood pressure, anemia, back pain in renal area, Darkness or redness in the urine etc. Therefore, the historical and diagnostic data of a patient when applied to data exploration and deep learning can help the physicians to predict the risk of early stage in CKD. According to the Global Data (2019) a leading data analytics company has specified that CKD had low diagnosis rate till 2026 unless there are effective ways incorporated to diagnose CKD at early stages. CKD progression cannot be analyzed by an instrument therefore physicians need a machine learning model which works on the clinical dataset of the patients to provide decisions in which type of patient to treat on time, progression of kidney failure stage in patients and ways to reduce the mortality rate[1]. Several research work has been done today in the field of chronic kidney failure prediction using machine learning. Hoerger et. al.(2005) developed a decision modeling combined with real-world dada and medical knowledge to predict CKD. Saurabh singh Thakur et. al.(2018) used non-contact sensor device data to monitor vital parameters like heart rate, respiration rate and heart rate variability of hemodialysis patients. He developed a Machine Learning based prediction model to predict CKD from sensor data. Jaya Sandeep (2018)[8] has developed the prediction model of CKD in patients using various classification algorithm through feature selection by Chi-Square test. The Random Forest classification technique provides 100% accuracy in prediction compared to other classifiers. Zeng et. al.(108) has specified that big data for CKD has vital opportunities and needs mature technology and policy framework to support and provide better health care through cross-disciplinary events. An automated Deep learning algorithm is evolved after comparing various classification model and its performance on the proposed dataset[6]. The paper is organized based on the following parts. Part I Describes the dataset proposed for the analysis, Part II performs the dataset validation and effectiveness of the sample size is predicted for classification performance using Data Augmentation and Learning curve model. Part III involves the technique used for data exploration and cleaning and Part IV involves feature engineering approach to select the appropriate features among 25 features involved in effective and efficient classification of CKD and non CKD. Part V incorporates the various classification model and its accuracy estimation. Part VI involves the proposed deep learning model and its performance analysis. The proposed model will help the nephrologist to predict the CKD status of the patients with fast accurate result. This leads to early treatment and reduces the speed of progression of the disease.

II. PROBLEM STATEMENT

Chronic kidney disease (CKD) is common. Kidney disease severity can be classified by estimated glomerular filtration rate (GFR) and albuminuria, but more accurate information regarding risk for progression to kidney failure is required for clinical decisions about testing, treatment, and referral. Design and implement a predictive analytical tool using Deep learning algorithm for efficient and accurate detection of severity of chronic kidney failure in patients in India and nature of medical treatment to be prescribed.

Revised Manuscript Received on July 06, 2019.

Dr. S. Pitchumani Angayarkanni, Associate Professor, Department of Computer Science, Lady Doak College, Madurai, TamilNadu, India.

III. DATASET PROPOSED FOR THE STUDY

The proposed dataset is taken from UCI Machine Learning Repository as specified in figure 1 below. The dataset was provided by Apollo hospital, Karaikudi, Tamil Nadu. The dataset consists of 400 instances with 25 attributes.

Fig 1. Dataset from UCI Repository

The proposed dataset has missing values and therefore it requires preprocessing.

IV. DATA PREPROCESSING

The data set consists of selected information from 400 patients. The 25 attributes with one dependent variable and 24 independent variables. The attributes involved are shown in figure 2

Attribute number	Attributes (type)	Attribute values	Attribute codes
1	Age (numerical)	Years	Age
2	Blood pressure (numerical)	mm/Hg	bp
3	Specific gravity (nominal)	1.005, 1.010, 1.015, 1.020, 1.025	sg
4	Albumin (nominal)	0, 1, 2, 3, 4, 5	al
5	Sugar (nominal)	0, 1, 2, 3, 4, 5	su
6	Red blood cells (nominal)	Normal, abnormal	rbc
7	Pus cell (nominal)	Normal, abnormal	pc
8	Pus cell clumps (nominal)	Present, not present	pcc
9	Bacteria (nominal)	Present, not present	ba
10	Blood glucose random (numerical)	mg/dl	bgr
11	Blood urea (numerical)	mg/dl	bu
12	Serum creatinine (numerical)	mg/dl	sc
13	Sodium (numerical)	mEq/L	sod
14	Potassium (numerical)	mEq/L	pot
15	Hemoglobin (numerical)	g	hemo
16	Packed cell volume (numerical)	-	pcv
17	White blood cell count (numerical)	cells/cumm	wbcc
18	Red blood cell count (numerical)	millions/cmm	rbcc
19	Hypertension (nominal)	No, yes	htn
20.	Diabetes mellitus (nominal)	No, yes	dm
21	Coronary artery disease (nominal)	No, yes	cad
22	Appetite (nominal)	Good, poor	appet
23	Pedal edema (nominal)	Yes, no	pe
24	Anemia (nominal)	Yes, no	ane
25	Class (nominal)	CKD, NOTCKD	-

CKD: Chronic kidney disease

Fig 2. Attributes involved in the dataset

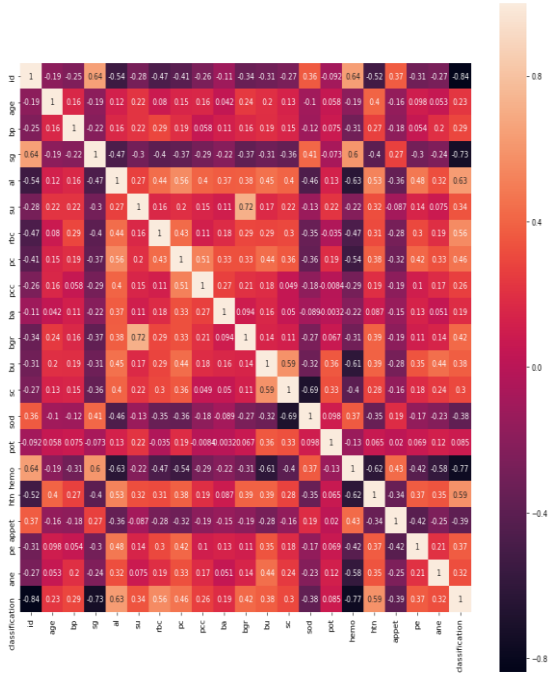

Fig 3. Heat Map of the dataset

The figure 3 and 4 indicates that the dataset has certain attributes with missing values. This clearly indicates that we need to perform data imputation to replace the missing values so that our prediction is valid and accurate. Missing values might lead to wrong prediction results. For categorical data type One-hot encoding is performed.

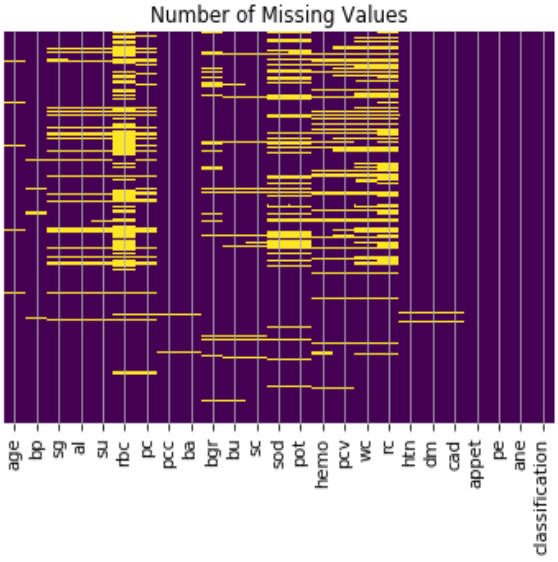

Fig 4. Distribution of missing values among the attributes

Since the data is found to be missing at random we apply Imputation methods as specified. Linear Regression approach is used to replace the missing values in the dataset for the attributes. This will yield unbiased estimate of the model parameter. Since the dataset involves missing at random data Imputation using Mean/ Median value id performed on the following attributes like age, hemo and pcv. We use a different way to impute is through matching: for each

unit with a missing, find a unit with similar values of Xin the observed data and take its Yvalue. This approach is also sometimes called “hot-deck” imputation. It is used for the following attributes like 'bp','sg','al','su','bgr','bu','sc','sod','pot', 'rc','wc' since theses are sensitive clinical attributes which plays a vital role in classification and prediction.

```
Data columns (total 26 columns):
id      233 non-null int64
age     233 non-null float64
bp      233 non-null float64
sg      233 non-null float64
al      233 non-null float64
su      233 non-null float64
rbc     233 non-null float64
pc      233 non-null float64
pcc     233 non-null float64
ba      233 non-null float64
bgr     233 non-null float64
bu      233 non-null float64
sc      233 non-null float64
sod     233 non-null float64
pot     233 non-null float64
```

Fig 5. Dataset after Preprocessing

Algorithm: Data Imputation

Input: Dataset D with missing value attributes like ‘age’,'bp','sg','al','su','hemo’,'bgr','bu','sc','sod','pot', 'rc','wc',’pcv’ 14 attributes.
Output: Dataset D with no missing values
Procedure:
For each missing value in the given dataset labelling:
Step 1: literation=1, impute=0
Step 2: Calculate Sign(i,j)
Step 3: If attributes found in important for clincical decision then
Perform Hot-Deck encoding
Else
Fill with the mean (for a continuous attribute) or mode (for a nominal attribute). Sort all Sign(i,j) Rank CIMV(i) in ascending order
Step 4 : Repeat the following two steps until convergence (k iterations).
Step 5: For i=1 to (Number of missing values) Impute CIMV/ Impute Hot-Dock encoding
(i) utilizing all the dataset based on the kNN algorithm literation ++; impute=1; Output: Results with filled-in values for all missing values.

Hot-deck(attributes)
1)Data set is divided as incomplete data set and complete data set.
2)Let X_i be the data matrix specifying the complete set of data, and x_{ij} be i^{th} observation pertaining to j^{th} variable. And let Y_i be a data matrix specifying the incomplete data set and y_{ij} be i^{th} observation pertaining to j^{th} variable.
3)Euclidian distances are computed for each line containing incomplete data set.

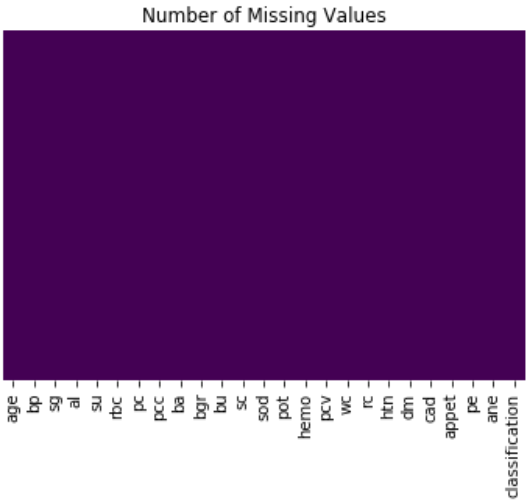

Fig 6. Shows the Pre-processed dataset with no missing values

The proposed imputation method performance is validated using the Standard error and correlation coefficient approach. Standard error is estimated as the difference between estimation observation and real observation.

$$\sigma = \frac{\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}}{\sqrt{n}} \qquad \text{Eqn. 1}$$

Where x_i is the estimated observation and y_i is the real observation as per the suggestions by the doctor.
Correlation coefficient is estimated between real and he observed value and specified as

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(\hat{X}_i - \overline{\hat{X}})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(\hat{X}_i - \overline{\hat{X}})^2}} \qquad \textbf{Eqn. 2}$$

Where Xi is the real value and $\hat{X}$ is the observed value.

Table 1. The performance evaluation

Attributes	Standard Error	Correlation Coefficient
14	0.01	0.984

The table 1 clearly indicates that the 14 attributes when missing values are replaced with the specified approach has a very low standard error and an effective correlation coefficient.

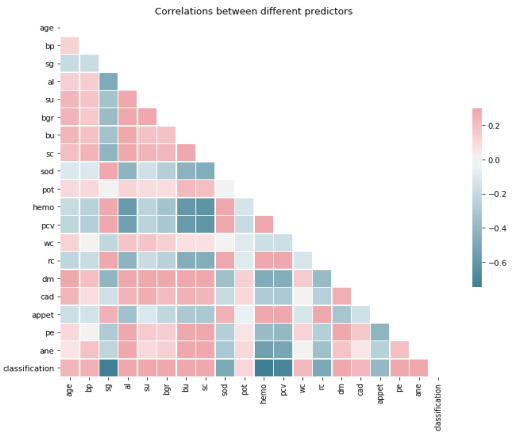

Fig 7. Correlation between the attributes

Figure 7 indicates the correlation between the attributes after replacing the missing values.

V. FEATURE ENGINEERING

Feature engineering refers to the process of using domain knowledge of the data to create features with which the machine learning algorithm works effectively. This is mainly used to remove the attributes which do not play a vital role in classification. The best features are chosen using a Wrapper method. Sequential **Feature Selection**. Sequential Forward **Selection** (SFS) method which is considered as a special case of WS is chosen. It is a greedy search algorithm which finds an optimal feature subset by iteratively selecting the features based on the classifier. The major benefits of feature selection include: Reduce Overfitting, Improve Accuracy and Reduce the training time.

Input: Set of features $X = \{x_0, x_1, \ldots, x_n\}$, size of feature set n, size of target feature subset d
Output: Suboptimum feature subset Y_{subopt} of size d
1: $Y_{subopt} \leftarrow \emptyset$
2: **for** $j = 1 \rightarrow d$ **do**
3: $x \leftarrow \max(J(Y_{subopt} \cap \{x_i\})) \mid x_i \in X \text{ and } x_i \notin Y_{subopt}$
4: $Y_{subopt} \leftarrow Y_{subopt} \cap \{x\}$
5: **end for**

Fig 8. Sequential Forward Selection Algorithm

Detailed classification report:

	precision	recall	f1-score	support
0.0	1.00	0.98	0.99	45
1.0	0.97	1.00	0.98	32
micro avg	0.99	0.99	0.99	77
macro avg	0.98	0.99	0.99	77
weighted avg	0.99	0.99	0.99	77

Fig 9. Confusion Matrix

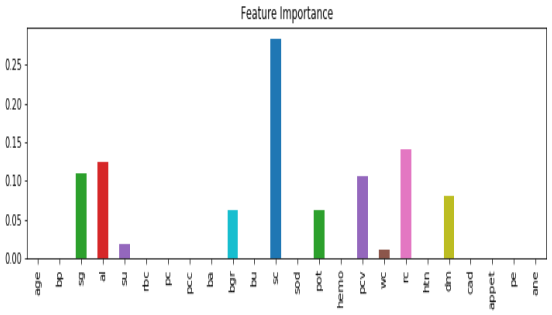

Fig 10. Features Extracted

The algorithm selects best 10 features out of 25 features which plays a vital role in classification of CKD Vs. non CKD. The predicted features are 'sg','al','su','bgr','sc','pot','pcv','wc','rc','dm'.

VI. COMPARISON OF VARIOUS CLASSIFICATION ALGORITHMS

The dataset after the feature selection is classified using eleven different classifiers to predict the accuracy of the data points used for classification and a learning curve is plotted to check the feasibility of the data points proposed for the analysis. The algorithms used for classification are depicted in table 2 with their functionalities. The Classification accuracy was determined by splitting the data points into training and test dataset at different cross folds. The cross folds of training and test samples include 20:80,40:60,50:50,60:40,80:20 of training and test dataset.
The Algorithm Selection approach is focused on:

- to determine which classifier performs best on a given dataset.
- predict this by training a meta-model on meta-data comprised of dataset characterizations, i.e., meta-features, and the performances of different classifiers on these datasets.

Table 2: Classification algorithms

Method	Explanation
K-Nearest Neighbors (KNN)	Classifying unknown examples by searching the closest data in pattern space.Prediction based on Euclidian Distance. The Euclidean distance $d(x,y)$ is used to measure the distance for finding the k closest examples in the pattern space. The class of the unknown example is identified by a majority voting from its neighbors.
Support Vector Machine (SVM)	The support vector machine (SVM) is the popular data mining method used to predict the category of data The main idea of SVM is to find the optimal hyperplane between data of two classes in the training data
Decision Tree (DT)	A decision tree is a structure that includes a root node, branches, and leaf nodes. It divides the data into classes based on the attribute value found in training sample.
Logistic Regression (LR)	Logistic Regression (LR) is the linear regression model. LR computes the distribution between the example X and boolean class label Y by $P(X\|Y)$.
Linear Discriminant Analysis(LDA)	The purpose of linear discriminant analysis (LDA) is to estimate the probability that a sample belongs to a specific class given the data sample itself.
Classification And Regression Trees(CART)	A node represents a single input variable (X) and a split point on that variable, assuming the variable is numeric. The leaf nodes (also called terminal nodes) of the tree contain an output variable (y) which is used to make a prediction.
Naïve Bayes (NB)	an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach.
Random Forest(RF)	Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.
Extreme Gradient Boosting(XGBoost)	XGBoost is one of the most popular machine learning algorithm these days. Boosting is a sequential technique which works on the principle of an ensemble. It combines a set of weak learners and delivers improved prediction accuracy. GBoost (Extreme Gradient Boosting) belongs to a family of boosting algorithms and uses the gradient boosting (GBM) framework at its core.

Gradient Boosting Regression(GBR)	**regression builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage, a regression tree is fit on the negative gradient of the given loss function.**
Ada Boosting (ABR)	**It combines multiple classifiers to increase the accuracy of classifiers. AdaBoost is an iterative ensemble method. AdaBoost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier. The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations. Any machine learning algorithm can be used as base classifier if it accepts weights on the training set.**

Before performing the proposed machine learning approach for classification of CKD it is necessary to validate the sample size of the data point to be used for classification. The Learning curve approach is used which is the (average) model performance as function of the training sample size[2].

- A partial learning curve is computed, using small samples, to measure how similarly algorithms behave on two data sets.
- Best on Sample method uses the performance estimates of classifiers on a small sample and recommends the classifiers which perform best on this sample, in descending order
- Design a study that evaluates model skill versus the size of the training dataset.
- Plotting the result as a line plot with training dataset size on the x-axis and model skill on the y-axis will give you an idea of how the size of the data affects the skill of the model on your specific problem.

This graph in figure 11 is called a learning curve
The accuracy of the proposed classifiers is depicted in figure 11 followed by the table 3 which specifies the accuracy of the data set under various cross folds using LR classifier.

Fig 11. Learning curve

The Learning curve converges when the training and test samples are chosen in the ratio of 40:60 and 50:50.

Table 3: Classifier accuracy

Training :Test dataset samples	LR	LDA	CART	XGB
20:80	0.97	0.95	1.0	1.0
40:60	0.98	0.96	1.0	1.0
60:40	0.97	0.95	1.0	1.0
80:20	0.95	0.90	1.0	1.0
100:100	1	1	1.0	1.0

The feasible data samples proposed for various classifiers is 40:60 ie. Out of 400 instances 160 data points were used as training dataset and the remaining 240 data points were used as test dataset. The accuracy of various classifier model applied on the 40:60 data samples with the 10 selected features are specified in figure 12.

```
Setup complete...
LR 1.0
LDA 0.961038961038961
KNN 0.7662337662337663
CART 0.987012987012987
NB 0.987012987012987
SVM 0.5844155844155844
RF 0.987012987012987
XGB 0.974025974025974
GBR 0.5844155844155844
ABR 0.987012987012987
```

Fig 12. Accuracy of 10 Classifier over 40:60 training and test data samples without feature selection

```
LR 0.9787234042553191
LDA 0.9787234042553191
KNN 0.8085106382978723
CART 1.0
NB 1.0
SVM 0.48936170212765956
RF 1.0
XGB 1.0
GBR 0.48936170212765956
ABR 1.0
```

Fig 13. Accuracy of 10 Classifier over 40:60 training and test data samples with feature selection

The figure clearly indicates that the accuracy of the classifier increases with feature selection approach.

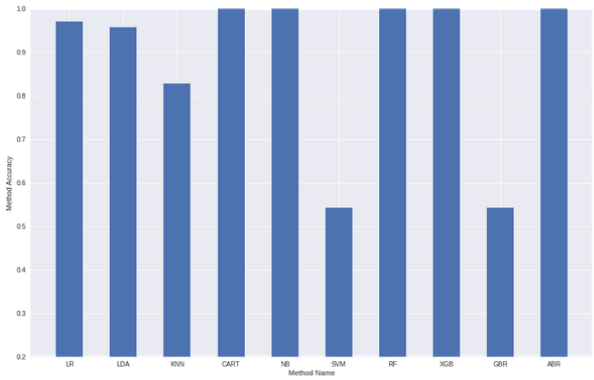

Fig 14. Accuracy of the proposed classifier

The figure 14 clearly indicates that LR,LDA have accuracy nearly 98% and CART,NB,RF,XGB and ABR has 100% accuracy in classification of the given data points into CKD and non CKD[4]. This clearly indicates that the sample size of the existing dataset is feasible to perform deep learning algorithm for predictive analytics.

VII. PROPOSED DEEP LEARNING MODEL

Recurrent Neural network is mainly used sequence prediction problem. It is classified into three types one-to-many, many to one and many to many. It is mainly used in classification prediction problem. It is linear architecture.

```
model = Sequential()
model.add(Dense(100, activation='relu', input_dim=26))
model.add(Dropout(0.5))
model.add(Dense(100, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(2, activation='softmax'))
model.compile(optimizer='nadam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])
model.fit(X_train,y_train,verbose=1,shuffle=True, nb_epoch=3,batch_size=100,validation_split=0.2

# Read test data and evaluate results


score = model.evaluate(X_test,y_test, batch_size=16)
print("LOSS")
print(score[0])
print("precision")
print(score[1])
```

```
10
Dataset size: 400
Epochs: 50
Learning rate: 0.001
Batch size: 40
Input dimension: 24
Features used: Index(['age', 'sg', 'al', 'bu', 'sc', 'sod', 'hemo', 'pcv', 'rc', 'dm'], dtype='object')
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_195 (Dense)            (None, 100)               1100
_________________________________________________________________
batch_normalization_37 (Batc (None, 100)               400
_________________________________________________________________
dense_196 (Dense)            (None, 50)                5050
_________________________________________________________________
batch_normalization_38 (Batc (None, 50)                200
_________________________________________________________________
dense_197 (Dense)            (None, 25)                1275
_________________________________________________________________
batch_normalization_39 (Batc (None, 25)                100
_________________________________________________________________
dense_198 (Dense)            (None, 1)                 26
_________________________________________________________________
```

Fig 15. Proposed Machine Learning Program

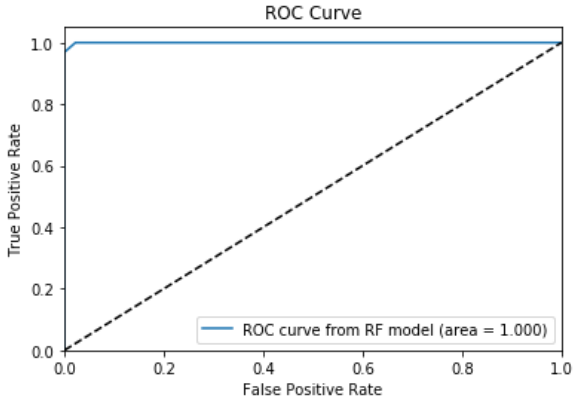

Fig 16.ROC

```
Epoch 97/100
1198/1198 [==============================] - 6s 5ms/step - loss:
0.0018
Epoch 98/100
1198/1198 [==============================] - 6s 5ms/step - loss:
0.0014
Epoch 99/100
1198/1198 [==============================] - 6s 5ms/step - loss:
0.0014
Epoch 100/100
1198/1198 [==============================] - 6s 5ms/step - loss:
0.0015
```

Fig 17. Epochs and MSE

Figures 15,16, and 17 clearly indicates the proposed RNN algorithm and the Receiver Operating Characteristic curve with an accuracy of 0.98 and minimum error of 0.0015. This clearly indicates that the proposed RNN algorithm perform accurate classification of the proposed dataset.

VIII. CONCLUSION

The proposed RNN algorithm helps in automatic detection of CKD with high accuracy of 98.99%. In future the same architecture can be used for analysis of various stages of Chronic Kidney Failure[7]. The Natural Language processing can be used in future to convert the medical records into electronic format and can be integrated with the proposed architecture to perform real time analysis of chronic

kidney failure. This can be coupled with the knowledge engine generated by the nephrologist opinion which can act as a decision support system to enhance the prediction of chronic kidney disease and the nature of treatment to be suggested in order to reduce the mortality rate of the patients.

REFERENCES

1. Sunethra Kanthi Gunatilake1*, S. Sunil Samaratunga2 and Ruvini Takshala Rubasinghe1,Chronic Kidney Disease (CKD) in Sri Lanka - Current Research Evidence Justification: A Review, Sabaragamuwa University Journal Volume 13 Number 2; December 2014.
2. Zeng, Xiao-Xi et al. "Big Data Research in Chronic Kidney Disease." *Chinese medical journal* vol. 131,22 (2018): 2647-2650. doi:10.4103/0366-6999.245275
3. Saran R, Robinson B, Abbott KC, Agodoa LY, Bhave N, Bragg-Gresham J, et al. US renal data system 2017 annual data report: Epidemiology of kidney disease in the United States. Am J Kidney Dis. 2018;71:A7. doi: 10.1053/j.ajkd.2018.01.002
4. Sossi Alaoui S., Aksasse B., Farhaoui Y. (2019) Statistical and Predictive Analytics of Chronic Kidney Disease. In: Ezziyyani M. (eds) Advanced Intelligent Systems for Sustainable Development (AI2SD'2018). AI2SD 2018. Advances in Intelligent Systems and Computing, vol 914. Springer, Cham
5. Greenberg JH, Kakajiwala A, Parikh CR, Furth S (2018) Emerging biomarkers of chronic kidney disease in children. Pediatr Nephrol 33:925–933
6. Aljaaf, Ahmed J. et al. "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics." 2018 IEEE Congress on Evolutionary Computation (CEC) (2018): 1-9.
7. Matsuzaki K, Suzuki H, Kobayashi T, Shimizu Y, Tomino Y (2016) Analysis of Predictive Factors for Deterioration of Renal Function in Chronic Kidney Disease Patients. J Nephrol Ther 6: 240. doi:10.4172/2161-0959.1000240.
8. Sandeep Reddy Mula, Jaya. (2018). Chronic Kidney Disease Analysis Using Machine Learning Algorithms. International Journal for Research in Applied Science and Engineering Technology. 6. 3367-3379. 10.22214/ijraset.2018.1469.
9. Zeng, X. X., Liu, J., Ma, L., & Fu, P. (2018). Big Data Research in Chronic Kidney Disease. Chinese medical journal, 131(22), 2647–2650. doi:10.4103/0366-6999.245275

AUTHORS PROFILE

Dr. S. Pitchumani Angayarkanni completed my Ph.D in 2017 from Karunya University, Coimbatore. Areas of Interest includes Medical Image processing, Data Analytics, Machine Learning and Natural Language Processing**.** Has more than 17 years of Teaching experience in computer science.