Iraqi Journal of Statistical Science (28) 2019
pp[140- 159]

Prediction and Factors Affecting of Chronic Kidney Disease Diagnosis using Artificial Neural Networks Model and Logistic Regression Model

Omar Qusay Alshebly
omarqusay@uomosul.edu.iq

Dr. Rizgar Maghdid Ahmed
rizgar.ahmed@su.edu.krd

Abstract

The last few years witnessed a great and increasing interest in the field of intelligent classification techniques which rely on Machine Learning. In recent times Machine Learning one of the areas in Artificial Intelligence has been widely used in order to assist medical experts and doctors in the prediction and diagnosis of different diseases. In this paper, we applied two different machine learning algorithms to a problem in the domain of medical diagnosis and analyzed their efficiency in prediction the results. The problem selected for the study is the diagnosis and factors affecting Chronic Kidney Disease. The dataset used for the study consists of 153 cases and 11 attributes of CKD patients. The objective of this research is to compare the performance of Artificial Neural Networks (ANNs) and Logistic Regression (LR) classifier on the basis of the following criteria: Accuracy, Sensitivity, Specificity, Prevalence, and Area under curve (ROC) for CKD prediction. From the experimental results, it is observed that the performance of ANNs classifier is better than the Logistic Regression model. With the accuracy of 84.44%, sensitivity of 84.21%, specificity of 84.61% and AUC_{ROC} of 84.41%. Also, through the final fitted models used, the most important factors that have a clear impact on chronic kidney disease patients are creatinine and urea.

Keywords:
Machine Learning, Logistic Regression, Artificial Neural Networks, Chronic Kidney Disease, Accuracy, AUC_{ROC}.

* Researcher / College of Computer science and Mathematics / Mosul University.
* Assit.Prof.Dr / College of Administration and Economic/ Salahaddin University.

Received Date:16/12 /2018 **Accepted Date: 2 /1 / 2019**

التنبؤ والعوامل المؤثرة في تشخيص مرض الفشل الكلوي المزمن باستخدام أُنموذج الشبكات العصبية الاصطناعية وأُنموذج الانحدار اللوجستي

الملخص

شهدت السنوات القليلة الماضية اهتمامًا كبيرًا ومتزايدًا في مجال تقنيات التصنيف الذكية التي تعتمد على تعلم الآلة. في الآونة الأخيرة ،تم استخدام التعلم الآلة وهي واحدة من مجالات الذكاء الاصطناعي على نطاق واسع من أجل مساعدة الخبراء الطبيين والأطباء في التنبؤ بأمراض مختلفة وتشخيصها. .في هذا البحث ، قمنا بتطبيق اثنين من مختلف خوارزميات التعلم الآلي للمشكلة في مجال التشخيص الطبي وتحليل كفاءتهما في التنبؤ بالنتائج. إنَ المشكلة المختارة للدراسة هي التشخيص والعوامل المؤثرة في مرض الكلى المزمن. تتكون مجموعة البيانات المستخدمة في الدراسة من 153 حالة و 11 سمة لمرضى الكلى المزمنة. إنً الهدف من هذا البحث هو مقارنة أداء الشبكات العصبية الاصطناعية والانحدار اللوجستي على أساس معايير الدقة والحساسية والخصوصية والانتشار والمنطقة تحت المنحنى للتنبؤ بمرض الكلى المزمن. من النتائج التجريبية لوحظ أن أداء مصنف الشبكات العصبية الاصطناعية أَفضل من أُنموذج الانحدار اللوجستي مع ما يقارب دقة 84.44% وحساسية 84.21% وخصوصية 84.61% والمساحة تحت المنحنى هي 84.41%.وكذلك من خلال النماذج النهائية القياسية فإن أهم العوامل المؤثرة التي لها تأثير واضح في دراستنا هذه في مرضى الكلى المزمنين هي الكرياتينين واليوريا.

1. Introduction

Last years, after increasing the number of patients with chronic kidney disease, It had to be highlighted to study this disease and the factors affecting it and the use of statistical methods and artificial intelligence techniques, artificial techniques have been receiving a lot of interest nowadays.

Chronic kidney disease (CKD) also known as chronic renal failure, or chronic kidney failure, is much more widespread than people realize; it often goes undetected and undiagnosed until the disease is well advanced. Often, CKD is diagnosed as a result of screening of people known to be at risk of problems, such as those with high blood pressure or diabetes and those who have relatives with CKD. It may also be identified when it leads to one of its recognized complications, such as cardiovascular disease or anemia. (Andrew et al., 2005) (Wang et al., 2013). The national estimated prevalence of CKD is 141 patients per million populations. Diabetes mellitus (33%) and HTN (22.6%) were the most common causes of CKD, followed by obstructive uropathy in 17.3%,

undetermined causes in 14%, pyelonephritis in 4.7%, glomerulonephritis in 4.3%, and polycystic kidney disease in 3.9% (Moukeh et al., 2009).

Machine learning is a branch of computational sciences that deals with learning the systems automatically based on inputs. The Classification is the main problem which is located in supervised machine learning.

Igor Kononenko (2001) presented a view on the use of Machine learning techniques 1) for the interpretation of medical data 2) for intelligent analysis of medical data in the current scenario and 3) for assistance of physicians in diagnosis of medical disorders, in the future. The authors suggested integration of machine learning techniques with the existing instrumentations for the acceptance of machine learning in medicine. Chen et al. (2009) presented A Comparative analysis of logistic regression, support vector machine and artificial neural network for the differential diagnosis of benign and malignant solid breast tumors by the use of three-dimensional power Doppler imaging. The diagnostic performances of these three models (LRA, SVM and NN) are not different as demonstrated by ROC curve analysis. In (2013) Abid Sarwar et al. compared the accuracy of Naïve Bayes, artificial neural network, and KNN algorithm for the type II diabetes. Type II diabetes is a condition in which the pancreas is not able to produce the needed amount of insulin or the cell is not able to use the produced insulin (insulin resistance) which leads to abnormal glucose level in the blood. The results showed that neural network with 96% prediction accuracy performs better than Naïve Bayes with 95% and KNN 91%.K.R.Lakshmi et al. (2014) proposed performance evaluation of three data mining techniques for predicting kidney dialysis survivability. In this research, various data mining techniques (Artificial Neural Networks, Decision tree and Logical Regression) are used to extract knowledge about the interaction between these variables and patient survival. A performance comparison of three data mining techniques is applied for extracting knowledge. Vijayarani et al. (2015) projected work on prediction of kidney disease using data mining classification algorithms. Prediction of four types of Kidney diseases namely Nephritic Syndrome, Chronic Kidney disease, Acute Renal Failure and Chronic Glomerulonephritis. Supervised classification algorithm Support Vector Machine (SVM) and Artificial Neural Network (ANN) is used to predict the kidney disease. Experimental results show that ANN is best classifier Classification accuracy for ANN is higher compared to SVM. Sharma et al.(2016) presented Different machine learning classification algorithm for diagnosis of chronic kidney disease is discussed. Various classification techniques that are used are: Decision Tree, Linear Discriminant

classifier, Quadratic Discriminant classifier, Linear SVM, Quadratic SVM, Fine KNN, Medium KNN, Cosine KNN, Cubic KNN, Weighted KNN, Feed Forward Back Propagation Neural Network using Gradient Descent and Feed Forward Back Propagation Neural.

The main aim of this study is to use two methods of supervised machine learning algorithms to predict and diagnose chronic kidney disease between two groups of patients (presence and absence) to identify the best classifier depend on a number of performance evaluation criteria. In addition, the study tested the most important factors affecting chronic kidney disease.

2. Methodology

2.1 Artificial Neural Networks

Scientists have found that the brain cortex contains 22 billion neurons and 220 trillion connections between them. This neuron that governs the mechanism of neuronal action drives neuroscientists, computer engineers and psychologists to try to simulate the human mind so that they can ultimately building a structure for giant computers that simulates the work of the human mind.

ANNs is a data processing system that simulates and resembles the way natural neural networks do to humans or to an organism. These elements relate to each other through a network of balanced links. The artificial neural network is an adaptive system, changing its structure based on the information through which it passes through the so-called learning stage (Wu and Larty, 2000).

On the other hand, ANNs has the right to solve many of the problems as it entered in many areas, the most important, field Medicine: An application of instant medicine which is related to the principle of memory as in the case of the human mind, principle of Pathological signs and diagnosis, the field of telecommunications: such as the disposal of resonance The sound that may be produced in the telephone lines, the military response to the target setting, and the field Banking: To open bank accounts by touch, sound or fingerprint Eye, as well as to identify bank signatures and handwriting. Business Areas: Networking Application In several businesses, especially in economic business(DA Silva et al., 2017).

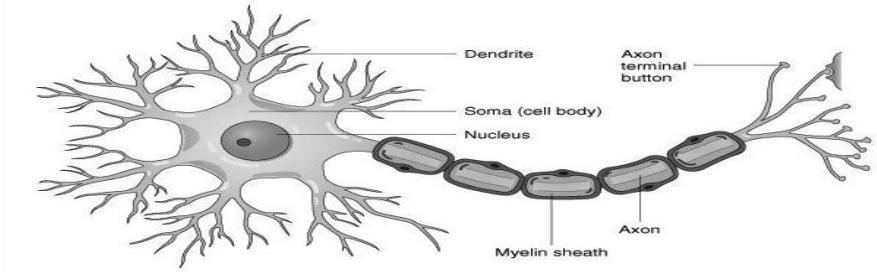

Figure (1) Biological Neural Network.

From Figure (1), the following can be clarified:

1-Dendrites: Entry points in each neuron which takes input from other neurons in the network in form of electrical impulses.

2-Cell Body (Soma): It generates inferences from the dendrite inputs and decides what action to take.

3-Axon terminals: They transmit outputs in form of electrical impulses to next neuron. (Graup, 2013).

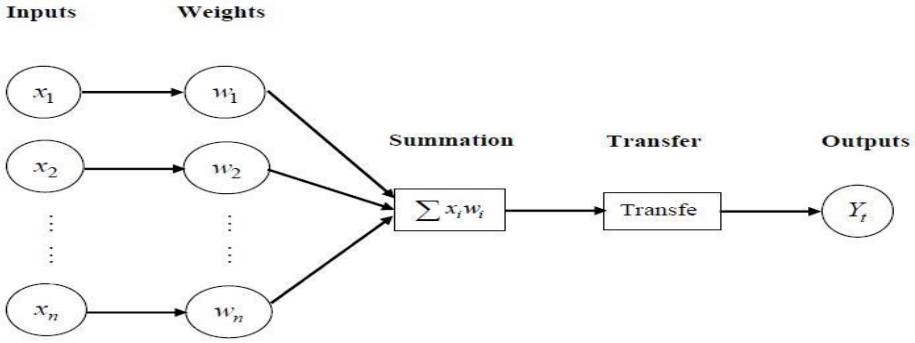

Figure (2) Artificial Neural Networks. It is clear from the previous figure 2 that the neural network consists of three segments as follows:

1-Input Layer includes values of x.

2-Hidden Layer includes the values of w and the resulting operations in the hidden slide.

3- Output Layer includes value of y.

The practical use of these networks lies in the possibility of applying algorithms designed to change the weight of the nodes connecting the artificial neurons together to produce a reaction (Wu and Larty, 2000).

$$net_{pj}^{L+1} = \sum_{i=1}^{n^L} w_{ij}^{L}\, out_i^L + bias_j^{L+1} \qquad (1)$$

$$out_{pi}^{L+1} = f\left(net_{pj}^{L+1}\right) = \frac{1}{1+e^{-\beta\, net_{pj}^{L+1}}} \qquad (2)$$

whereas:-

net_{pj}^{L+1}: Total entries multiplied by weights for unit j in layer (L + 1).

out_{pi}^{L+1}: The output of the layer (L + 1).

w_{ij}^{L}: Weights of layer L.

$bias_j^{L+1}$: Error value between layers.

2.2 Artificial Neural Network Learning Algorithms

The weights in the artificial neural network represent the initial information that the network will learn. Therefore, the weights must be updated during the training phase. So, several algorithms are used and depending on the type of network for this update in weights. In this syudy, we used the Backpropagation Algorithm, which is used in multi-layered and nonlinear neural networks. It is implemented in two main stages: (Livingstone, 2008)

1- Feed Forward Propagation.

2- Back Propagation.

First: Feed Forward Propagation

At this stage, no weight adjustment is made, but the network begins by assigning each processing element from the input layer to the excitation of the units of this layer, followed by a forward spread of that excitation across the rest of the grid layers. That is, the output of any layer affects only the next layer, and there is no correlation between the cells of the single layer. (Livingstone, 2008).

Second: BackPropagation

Is the stage at which the weights are set which the standard back-propagation algorithm in the network is the Gradient Descent algorithm, which allows the signal to be re-exported from the output to the input in reverse. The network weights are set by calculating the error between the output and the target. The algorithm can be represented for one repetition (Reed and Marks, 1999):

$$W_{k+1} = W_k - \alpha_k \cdot g_k \qquad (3)$$

Whereas:

W_k*:* weights vector.

α_k*:* learning rate.

g_k*:* current slope.

2.3 Logistic Regression Model

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. Quite often the outcome

variable is discrete, taking on two or more possible values. The logistic regression model is the most frequently used regression model for the analysis of these data. (David et al., 2013)

LR defined was known to be a type of nonlinear regression model describe The relationship between the dependent variable (the response) and a set of explanatory variables is determined A nonlinear relationship, where the dependent variable (response) is variable qualitative. The dependent variable in the regression model may take only two forms and syllabus (0) or (1), which is the basis for the Binary logistic regression, which was studied in our research. Alternatively, the variable may take more than two classes, which is called a multinomial logistic regression as for the explanatory variables; these variables can be continuous or discrete. (Harrell, 2010).

The logistic regression model is based on a basic assumption that dependent variable to be studied is a two-character variable and follows a Bernoulli distribution according to the probability function known as the following formula: (Özkale and Arıcan,2016)

$$p(Y = y_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} \qquad (4)$$

y_i =0 or 1.

The probability (π_i) can be defined mathematically in terms of explanatory variables and the logistic function as in the following formula:-

$$\pi(x_i) = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}} \qquad (5)$$

where (β:) Vector of parameters and $X_i = \{1, x_{i1}, x_{i2}, \dots x_{ip}\}$: Row vector of independent variables.

In order to simplify notation, we use the quantity π(x)=E(Y/X) to represent the conditional mean of Y given X when the logistic distribution is used. The specific from of the logistic regression model we use is:

$$\text{logit}\,(\pi(x_i)) = \ln \frac{\frac{e^{X_i\beta}}{1+e^{X_i\beta}}}{1-\frac{e^{X_i\beta}}{1+e^{X_i\beta}}} = \ln \frac{\frac{e^{X_i\beta}}{1+e^{X_i\beta}}}{\frac{1}{1+e^{X_i\beta}}} = \ln\left(e^{X_i\beta}\right) = X_i\beta = \left(Bo + \sum_{j=1}^{p} Bj\ xij\right) \quad (6)$$

Whereas:

$\beta_0, \beta_1, \dots, \beta_p$: Unknown parameters were estimated.

X_{ij}: Independent variables.

2.4 Performance Evaluation

In this part we will discuss several methods of evaluating the performance of (LR and ANNs).

A- Confusion Matrix

The classification matrix is a statistical indicator of the suitability of the model and thus its compatibility with the data. where it works on classification of binary events by using the Confusion Matrix, which shows the actual versus predicted affiliation of each group .(Soderstorm and Leitner, 1997)

Table (1): Confusion Matrix

Classification		**observation**	
		Negative	**Positive**
Observations	**Negative**	**True negative (TN)**	**False positive (FP)**
	Positive	**False negative (FN)**	**True positive (TP)**

B- Accuracy

Accuracy is the measure of how good our model is. It is expected to be closer to 1, if our model is performing well.

Accuracy = (TP+TN)/N (7)

Whereas:-

TN: The number of samples classified as negative (does not have the characteristic) is actually negative.

TP: The number of samples classified as positive (possessing the characteristic) is in fact positive.

N: Total number of samples (Soderstorm and Leitner, 1997).

B- Sensitivity

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (8)$$

Where FN the number of samples classified as negative is actually positive, and the model becomes more sensitive if he can identify the largest number of positive cases (Soderstorm and Leitner, 1997).

C- Specificity

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (9)$$

Where FP the number of samples classified as positive is actually negative, and the model becomes more specific if it can determine the largest number of negative cases (Soderstorm and Leitner, 1997)

D- Prevalence

The percentage of a population that is affected with a particular disease at a given time**.**

Prevalence = $\frac{TP+FN}{N}$ (10)

E- The Area under curve (ROC curve)

AUC is defined as a measure for the overall performance of the classifier scores across all possible values of the threshold (or cutoff_ point).

If the probability distributions are known for both detection and false alarms, it is possible to create a ROC curve by plotting the cumulative distribution (The area under probability from (-∞ to+∞), usually area under curve ROC using as a measure of the quality of probability classification. The area under curve used the following formula. (Hosmer & Lemshow,2000)

$$A_{ROC} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} = \frac{1}{PN}\int_0^N TP * dFP \quad (11)$$

3. Results and Discussion

3.1 Real Dataset Collection

This data contains 153 patients, including 12 variables, 11 of which are independent variables and a dependent variable (presence 1 and absence 0) of chronic kidney disease (CKD) depend on Blood test(serum), there is no missing value in this data. The patients included in this study were collected from Erbil teaching hospital during the period from first of April, 2018 to 30th June 2018.

The studied samples consist of 85 absence CKD and 68 presence CKD patients, the age of patients ranged from 12 to 90 years, also studied consist of 83 (54%) males and 70 (46%) females.

Table 2 shows the description of the data studied in this paper

Table (2): Description of the study variables

No.	Variable Name	Coding of Variable	Types of Variable
1.	Class(y)	1 presence of CKD. 0 absence of CKD.	nominal
2.	Sex(x_1)	1 male. 2 female.	nominal
3.	Age(x_2)	NA*	numerical

No.	Variable Name	Coding of Variable	Types of Variable
4.	Smoking(x_3)	1 smoked 2 non smoked	nominal
5.	Urea(x_4)	N/A	numerical
6.	Creatinine(x_5)	N/A	numerical
7.	Calcium(x_6)	N/A	numerical
8.	Phosphorus(x_7)	N/A	numerical
9.	Alkaline phosphateas(x_8)	N/A	numerical
10.	Glucose(x_9)	N/A	numerical
11.	Albumin(x_{10})	N/A	numerical
12.	total Bilirubin(x_{11})	N/A	numerical

* Not Available

Dataset is randomly partitioned into the training dataset and the test dataset, where (70%) (108 patients) of the samples are selected for training dataset and the rest (30%) (45 patients) are selected for the testing dataset. For a fair comparison and for alleviating the effect of the data partition, all the used classification methods are evaluated, for theirclassification performance metrics using 10 folds cross-validation, averaged over 10 partitioned times. All the implementations of the study on real data applications are carried out using R version (3.4.4).

3.2 Performance Evaluation of Models Applied

After partitioned the data into two groups (training and testing) we begun building the model based on the training dataset which includes 108 cases.

(0=65 cases 1=43 cases)

Now, 0 mean absence CKD, 1 mean presence CKD. The content of the r esponse is unchanged: 65 cases of absence class and 43 cases of presence class were detected.

Table (3) shows evaluation criteria for the LR model of the training da taset.

Table (3): Confusion Matrix and Statistics for Training Dataset of LR

Classification	**Prediction**	
	Absence 0	**Presence 1**

Observations	**Absence 0**	62	8
	Presence 1	3	35
Model Accuracy		**89.81%**	
Model Sensitivity		**81.40%**	
Model Specificity		**95.38%**	
Prevalence		**39.81%**	

From table 3 we can conclude that the logistic regression model has been able to properly classify 97 cases out of 108 available. the model has Accuracy (89.81), but also the sensitivity and the specificity are greater than 80 percent (81.40%, and 95.38%) respectively, That is, the model can predict correctly based on the independent variables entered by 81.40% for those with CKD patients. The specificity of the model was 95.38%, that is to say, it can predict correctly based on the independent variables entered by 95.38% for those without CKD. In addition, the prevalence of the disease in the community for this model of the training dataset is 39.81%.

Table (4) shows evaluation criteria for the LR model of the testing dataset.

Table (4): Confusion Matrix and Statistics for Testing Dataset of LR

Classification		**Prediction**	
		Absence 0	**Presence 1**
Observations	**Absence 0**	21	3
	Presence 1	5	16
Model Accuracy		**82.22%**	
Model Sensitivity		**84.21%**	
Model Specificity		**80.77%**	
Prevalence		**42.22%**	

From table 4 we found that all values have been dropped from the table 3 where the testing dataset has been achieved the model has Accuracy (82.22%), but also the sensitivity and the specificity are greater than 80 percent (84.21%, and 80.77.as well as, we also found that the prevalence of the disease in the population for this model of the testing dataset is 42.22%.

Another tool to measure the model performance is the Receiver Operator Characteristic (ROC). It determines the model's accuracy using area under curve $(AUC)_{ROC}$. the value $(AUC)_{ROC}$. for testing dataset of LR model is (0.8249) , As shown in Figure 3.

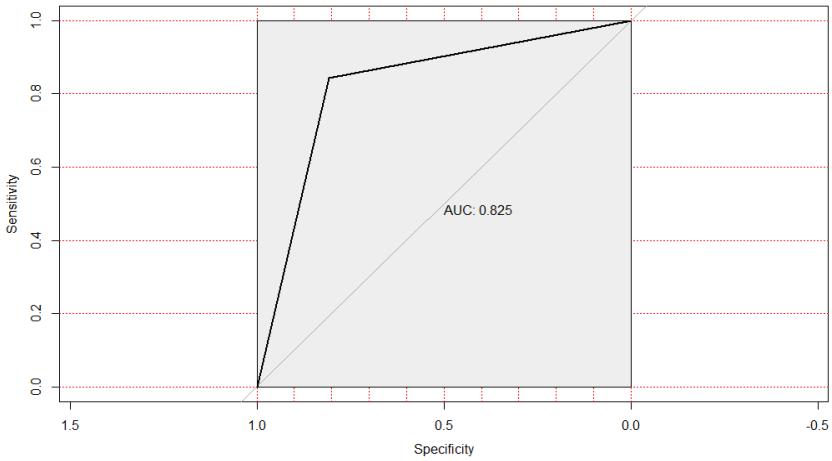

Figure (3) (ROC) curve of Testing dataset for LR

ROC is plotted between the sensitivity (y axis) and the specificity (x axis).from figure 3 shows area under curve value is 82.5%, The ROC is a metric used to check the quality of classifiers.

Table (5) shows the classification table and evaluation criteria for the Artificial Neural Networks model of the training dataset.

Table (5): Confusion Matrix and Statistics for Training Dataset of ANNs

Classification		**Prediction**	
		Absence 0	**Presence 1**
Observations	**Absence 0**	**65**	**0**
	Presence 1	**0**	**43**
Model Accuracy		**100%**	
Model Sensitivity		**100%**	
Model Specificity		**100%**	
Prevalence		**39.81%**	

Table 5 summarizes, the analysis of the confusion matrix, we can see that the Neural Networks model has been able to properly classify 108 cases out of 108 available. without any classification errors. As it is possible to verify, the model has Accuracy (100%), this indicates that all

other metrics (model sensitivity, model specificity, etc.) will be equal to the correct one hundred (100%).In addition; we also found that the prevalence of the disease in the community for this model of the training dataset is 39.81%.

From the previous results we found that the models of artificial neural networks have achieved excellent results for training dataset compared with the logistic regression model.

Table (6): Confusion Matrix and Statistics for Testing Dataset of ANNs

Classification		**Prediction**	
		Absence 0	**Presence 1**
Observations	**Absence 0**	22	3
	Presence 1	4	16
Model Accuracy		**84.44%**	
Model Sensitivity		**84.21%**	
Model Specificity		**84.61%**	
Prevalence		**42%**	

From table 6 we found that all values have been decreased from the table 5 where the testing dataset has been achieved the model has Accuracy (84.44%), but also the sensitivity and the specificity are greater than 80 percent (84.21%, and 84.61%)respectively. That is, the model can predict correctly based on the independent variables entered by 84.21% for those with CKD patients. The specificity of the model was 84.61%, that is to say, it can predict correctly based on the independent variables entered by 84.61% for those without CKD. As well as, we also found that the prevalence of the disease in the community for this model of the testing dataset is 42%.

On the other hand Area under the curve $(AUC)_{ROC}$: 0.8441296 shows in figure 4.

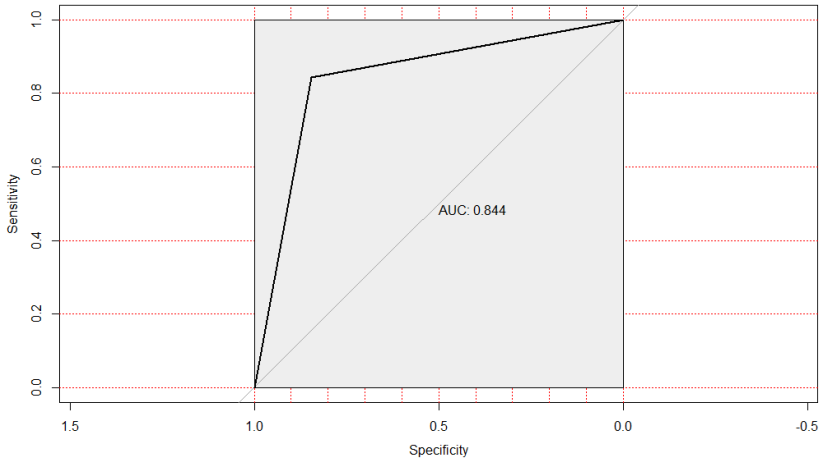

Figure (4) (ROC) curve of Testing dataset for ANNs

Figure (4) shows the ROC curve of area under the ROC curve was 84.4%, the area under curve was higher than 50%.

In the other word, we discussed the Comparison between two classification models (LR and ANNs) based on criteria (Model Accuracy, Model Sensitivity, Model Specificity, Prevalence, and Area under curve (ROC)).

Table 7 shows comparison for testing dataset only. Because the best classifier based on testing dataset. The results showed that ANNs model was better than LR where the classification using ANNs model was more accurate and more efficient.

Table (7): Performance Evaluation Criteria between Models

Model	Logistic Regression	Artificial Neural Networks
Model Accuracy	82.22%	84.44%
Model Sensitivity	84.21%	84.21%
Model Specificity	80.77%	84.61%
Prevalence	42.22%	42%
$(AUC)_{ROC}$	82.49%	84.41%

Table 7 summarizes, the accuracy of LR was 82.22%. While the accuracy of ANNs model was equal to 84.44%. This gives the preference for ANNs according to the accuracy of the model mean more accurate the model, then best model. The model sensitivity criterion for LR was 84.21%. While the model sensitivity criterion for an ANNs model was equal to 84.21%. It is a same value. In addition, the model Specificity

criterion for LR was 80.77%. The performance of the model was improved by using ANNs. The value was 84.61%. Means that last model has a complete preference.as well as that prevalence of the disease in the community all models have their value approximately 42%.

On the other hand, the area under curve (ROC) for the logistic regression was 82.49%. while area under curve (ROC) criterion for ANNs models was equal to 84.41%. The greater the value of area under curve (ROC), It was the best.

3.3 Fitted Final Model and Variables Importance

When dealing with neural network models, the study data is introduced to the neural network model and uses a single-hidden layer consisting of eight neurons in our study, as well as these have made many changes to reach the result above which follows a lot of changed in layers and nodes within each layer.by using Garson's algorithm (Garson, 1991) to evaluate relative variable importance (Beck, 2018).

We found importance variables for ANNs model which appeared in table 8 and figure 5.

Table (8): Variables Importance for ANNs

Garson algorithm for variable importance	
	Overall
creatinine	0.40
urea	0.18
total..Bilirubin	0.09
age	0.08
sex	0.07
Alk.phosphatas	0.05
phosphorus	0.04
Glucose	0.04
Albumin	0.03
calcium	0.02
smoking	0.00

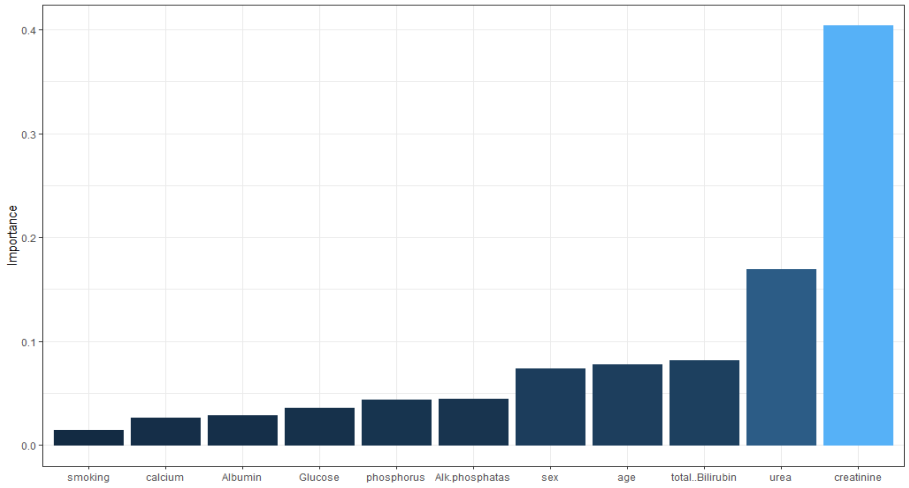

Figure (5) Importance Variables ANNs Model plot

Table 8 and figure 5 shows the degree of importance for each factor affecting CKD patients using ANNs, where the factor of creatinine was the most influential factor by 40% on the dependent variable (class), and urea factor at 18%, total bilirubin 9%, age of patient 8% also sex of patient 7%.

The following factors were the least affected factors in the dependent variable, ie, phosphorus 4%, glucose 4% and calcium 2%.

On the other hand, to determine the most important factors affecting the model table 9 shows importance variables for LR model used in our study. By using VarImp function in the caret package of R.

Table 9 shows significance of independent variables of LR model.

Table (9): Variables Importance for LR

glm variable importance	
	Importance
creatinine	100.00
urea	92.92
smoking	58.33
total..Bilirubin	46.07
Albumin	44.76
phosphorus	39.86
Glucose	37.43
sex	16.18
age	10.71
Alk.phosphatas	10.44
calcium	0.00

Table 9 shows the degree of importance for each independent variable affecting CKD Using LR, Where the variable creatinine is the largest effect followed by variable urea (100%, 95%) respectively and then smoking etc.

Figure 6 shows significance of independent variables of logistic regression model.

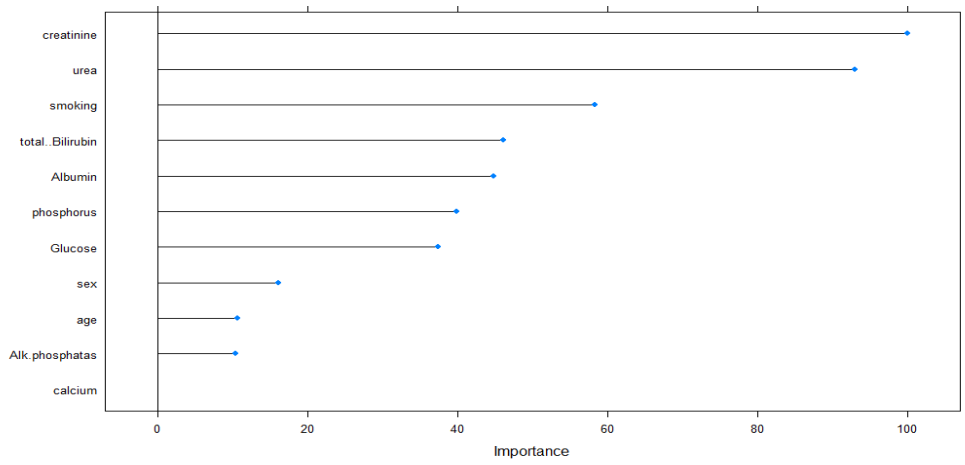

Figure (6) Importance Variables LR Model plot

From the results above may be determined the equation of logistic regression model with significant independent variables affecting of CKD patients. As shown in Table 10.

Table (10): Factors Affecting of CKD

Variable	Estimate β	Std. Error	Signi.	Wald value	df	p-value
urea	0.074279	0.033435	0.0263 *	4.935363	1	0.028662*
creatinine	6.981951	2.940306	0.0176 *	5.638564	1	0.019559*

* Significant if (p-value ≤ 0.05).

$$log\ odds = (-10.40) + 0.074x_1 + 6.98x_2$$

Whereas x_1:- urea independent variable. and x_2:- creatinine independent variable.

4. Conclusions

In view of great importance of CKD and what may be caused of death

and healthy crisis for community. The study has reached through comparison between of two classifier methods in the classification of CKD patients relying on the blood test, and based on evaluation criteria that the method of ANNs is the best methods used in this study. As well as we concluded the factors had greatest effect on the data of patients with chronic kidney disease that the variables creatinine and urea are the most effective and significant variables by using the two methods (LR and ANNs).

The study also concluded from the sample taken of the community, that the prevalence of disease at community has a big percent. Calls for action to reduce prevalence among the population.

5. References

1. Abid Sarwar, Vinod Sharma,(2013) ."Comparative analysis of machine learning techniques in prognosis of type II diabetes". AI & Society, Springer Verlag .
2. Andrew SL, Kai UE, Yusuke T, Adeera L, Josef C, Jerome R et al.(2005). "Definition and classification of chronic kidney disease: A position statement from kidney disease: Improving Global outcomes(KDIGO)Kidney Inter"; 67:2089-2100.
3. Beck, M. (2018)." Neural NetTools: Visualization and Analysis Tools for Neural Networks".
Chen, S.-T., Hsiao, Y.-H., Huang, Y.-L., Kuo, S.-J., Tseng, H.-S., Wu, H.-K. & Chen, D.-R. (2009). "Comparative analysis of logistic regression, support vector machine and artificial neural network for the

4. differential diagnosis of benign and malignant solid breast tumors by the use of three-dimensional power Doppler imaging." Korean Journal of Radiology, 10, 464-471.
5. Cho, S.-B. & Won, H.-H. (2003)."Machine learning in DNA microarray analysis for cancer classification". Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics, 19, 189-198, 2003. Australian Computer Society, Inc., 189-198.
6. Da Silva, I. N., Spatti, D. H., Flauzino, R. A., Liboni, L. H. B. & Dos Reis Alves, S. F. (2017). "Artificial neural networks". Cham: Springer International Publishing.
7. David W. Hosmer, Jr., Stanley Lemeshow and Rodney X. Sturdivant (2013)."Applied Logistic Regression", 3rd Edition. , John Wiley & Sons.
8. Garson, G.D. (1991). "Interpreting neural network connection weights". Artificial Intelligence Expert. 6(4):46–51.
9. Graupe, D. (2013). "Principles of artificial neural networks", World

Scientific.
10. Harrell, F. (2010)." Regression Modeling Strategies: With Applications to Linear Models", Logistic Regression and Survival Analysis. New York, Springer-Verlag.
11. Hosmer, D. and Lemeshow, S. (2000). Applied Logistic Regression. 2nd edition. New York: Johnson Wiley & Sons, Inc .
12. Igor Kononenko, (2001)."Machine Learning for Medical Diagnosis: History, State of the Art and perspective". Artificial Intelligence in Medicine, Elsevier, Vol.23, No. 1.
13. Inan, D., & Erdogan, B. E. (2013)."Liu-type logistic estimator, Communications in Statistics-Simulation and Computation", 42(7)pp.1578-1586.
14. K.R.Lakshmi, Y.Nagesh and M VeeraKrishna, (2014)."Performance comparison of three data mining techniques for predicting kidney disease survivability", International Journal of Advances in Engineering & Technology.
15. Livingstone, D. J. (2008). "Artificial neural networks: Methods and applications (methods in molecular biology)", Humana Press.
16. Moukeh G, Yacoub R, Fahdi F, Samer R, Sami A (2009)." Epidemiology of hemodialysis patients in Aleppo city". Saudi J Kidney Dis Transpl.; 20(1): 140-146.
17. Özkale, M. R., & Arıcan, E, (2016)."A new biased estimator in logistic regression model. Statistics", 50(2),pp. 233-253.
18. Priddy, K. L. & Keller, P. E. (2005). "Artificial neural networks: an introduction", SPIE press.
19. Reed, R. and Marks, R. J., (1999) ."Neural Smithing : Supervised Learning in Feedforward Artificial Neural Networks", Massachusetts Institute of Technology, England.
20. Sharma, S., Sharma, V., & Sharma, A. (2016). "Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis". arXiv preprint arXiv:1606.09581.
21. Soderstrom, R. and Leitner ،W. (1997). "The Effects of Base Rate, Selection Ratio،Sample Size ،and Reliability of Predictors on Predictive Efficiency Indicies Associated with Logistic Regression Models". Paper Presented at the Annual Meeting of the Mid- Western Educational Research Association (Chicago, IL, October 15-18,1997).
22. Vijayarani, S., Dhayanand, M. S., & Phil, M. (2015)."Kidney disease prediction using svm and ann algorithms". International Journal of Computing and Business Research (IJCBR) ISSN (Online), 2229-6166.
23. Wang C, CuiCui L, Gong W, Tanqi L (2013). "New urinary biomarkers for diabetic kidney disease". Biomarker Res; 1(9):1-4.
24. Wu, C.H. and Larty, J.W., (2000)."Neural Networks and Genome

Informatics", Elsevier Science Ltd. All rights reserved First edition, Amsterdam, USA.
Yegnanarayana, B. (2009). "Artificial neural networks", PHI Learning Pvt. Ltd.