Journal of Engineering and Applied Sciences 14 (15): 5167-5175, 2019
ISSN: 1816-949X


# Performance Evaluation of Diagnosis Chronic Kidney Disease using Support Vector Machine and Logistic Regression Model

[1]Rizgar Maghdid Ahmed and [2]Omar Qusay Alshebly
[1]Department of Statistics and Informatics, College of Administration and Economic,
Salahaddin University, Erbil, Iraq
rizgar.ahmed@su.edu.krd
[2]Department of Statistics and Informatics, College of Computer Science and Mathematics,
Mosul University, Mosul, Iraq
omarqusay.90@gmail.com

**Abstract:** With the rapid development of intelligent classification techniques which depends on machine learning, this study addressed the comparison between one of the traditional statistical models (logistic regression) with the supervised machine learning model (support vector machine) in order to classify chronic kidney disease patients based on a blood test (serum) for a group of presence and absence patients. The dataset contains data of 153 cases and 11 attributes for diagnosis of chronic kidney disease. The dataset were divided into two groups (training and testing) and after applied the above models depend on evaluation performance criteria (model accuracy, model sensitivity, model specificity, prevalence, kappa coefficient and area under curve (ROC)). The study concluded the results indicate SVM Model is the best performer (best classifier). As well the study concluded through the final fitted models used that the most important factors that have a clear impact on chronic kidney disease patients are creatinine and urea.

**Key words:** Classification, logistic regression, support vector machine, chronic kidney disease, accuracy, kappa coefficient area under curve (ROC)

## INTRODUCTION

In recent years, after increase the incidence of chronic kidney disease, it was necessary 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 now a days.

Chronic Kidney Disease (CKD) describes the gradual loss of kidney function. Your kidneys filter wastes and excess fluids from your blood which are then excreted in your urine. When chronic kidney disease reaches an advanced stage, dangerous levels of fluid, electrolytes and wastes can build up in your body (Jha *et al*., 2013). The reported prevalence of CKD in the NHANES (National Health and Nutrition Examination Survey) between 1999 and 2006 was 26 million out of a population base of approximately 200 million in the United States, of these, 65.3% had CKD, those with diabetes and hypertension had far greater prevalence of CKD (37 and 26%), respectively, compared to those without these conditions approximately 11%. 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. Classification models predict class labels for objects. Often, evaluation performance of classification models depends on results of training and testing sets. A training set is a set of data used to learn (building) a classifier while testing set of data is used to assess the strength classifier.

SVM is one of the good scientific methods for the classification algorithms which it is a tool for machine learning. SVM considered the most appropriate algorithm in machine learning. Practical experiments proved superior SVM on some older classification algorithms in many drawbacks (Byvatov *et al*., 2003; Colas and Brazdil, 2006).

The use of logistic regression has evolved in the study of the effect of the relationship between a set of independent variables on the response variable (depended variable) when it is categorical variable. In turn, LR is divided into two parts: the first binary logistic regression model consisting of only two classes. The second multinomial logistic regression depended variable which ordinal or nominal, it consist more two classes. Many researchers used LR Model, the first to use logistic function Verhulst 1920, who studied the relationship of

**Corresponding Author:** Rizgar Maghdid Ahmed, Department of Statistics and Informatics,
College of Administration and Economic, Salahaddin University, Erbil, Iraq, rizgar.ahmed@su.edu.krd

this function to population growth (Inan and Erdogan, 2013). The goal of classification is to learn a model that can accurately predict the class labels of new unseen test instances such as cancer prediction diagnostic problem (Cho and Won, 2003). The strength of classifier, here, appeared after the medical diagnosis for patient and will be the most powerful classification model which has the highest accuracy model, sensitivity model, specificity model, prevalence, kappa coefficient and area under curve (ROC).

**Literature review:** During the past few years, the number of classification techniques has increased with the rapid growth of technology, there are many researchers interested in this subject.

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. Depending on user emphasis for the use of ROC curve findings, the use of LRA appears to provide better sensitivity as compared to the other statistical models. By Bhatla and Jyoti (2012) heart disease prediction is done using three data mining techniques namely neural network, decision tree and Naive Bayes. Their results disclose that neural networks with 15 features have surpassed two other techniques and accordingly are selected as the predictive mode.

Sarwar and Sharma (2014) compared the accuracy of Naive Bayes, artificial neural network and kNN algorithm for the type 2 diabetes. Type 2 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 Naive Bayes with 95% and kNN 91%.

George *et al.* (2014) presented a diagnosis system for breast cancer using different machine learning algorithms such as support vector machines and neural networks they report accuracy rates which ranged from 76-94% on a dataset of 92 images.

Vijayarani and Dhayanand (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.

## MATERIALS AND METHODS

**Support vector machine:** The field of machine learning that cares with problems of classification is titled "supervised learning". Supervised learning points to the case where the researcher adopts a model using training data and then tests the model on a test data to see how well it does at portending class membership. One of the most famous tools of machine learning is SVM.

Support Vector Machines (SVM) is learning systems that use a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that tool a learning bias derived from statistical learning theory. This learning strategy introduced by Vapnik and co-workers is a principled and very powerful method that in the few years, since, its introduction has already outperformed most other systems in a wide variety of applications (Cristianini and Shawe-Taylor, 2000).

In general, there are two main types of SVM: linearly separable and nonlinearly separable. Firstly, the linearly separable (hard-margin) SVM is talked where the training data are linearly separable in input space generalized by Boser *et al.* (1992). Secondly, Cortes and Vapnik (1995) proposed nonlinearly separable (Soft-margin) SVM which allowed classification in the case of overlapped data.

**Linear separable support vector machine:** If we have a two-class classification problem, given training set containing n input vectors $x_i$ from d-dimensional input space and $y_i$ is the class label. The hard-margin SVM which separates the data points using a linear decision boundary where the function is a linear decision function defined as (Tan *et al.*, 2005):

$$f(x) = w^T + b = 0 \quad (1)$$

Whereas, w is a d-dimensional weight vector and b is a Bias term. And by rewriting the values of and the equations of the two supporting hyper-planes can be defined as:

$$S_1 = w^T X_{S1} + b = 1 \quad (2)$$

$$S_2 = w^T X_{S2} + b = -1 \quad (3)$$

with a normal vector w, since, these two hyper-planes are parallel and have the same normal vector. And after

making several derivations. Thus, the classification of unknown samples like depends on them and can be explained as:

$$d(x) = \text{sign}(\sum_{i=1}^{n} \alpha_i y_i x_i^T x + b) = \text{sign}(\sum_{SV=1}^{\#SV} \alpha_i y_i x_i^T x + b) \quad (4)$$

The Eq. 4 can be used to classify any unknown sample to a positive or negative class.

**Nonlinear separable support vector machines:** Many applications of life cannot be separated linearly in the input space except through the use of nonlinear techniques and it is known that the cost of nonlinear calculations is greater than the cost of linear calculations, therefore, there are two suited main approaches for solving these problems inner product process (kernel trick) and soft-margin SVM:

**Kernel trick:** One of the most important approaches used in machine learning includes changing the representation of the data. Burges (1998) introducing a technique known as the "kernel trick", kernel trick is used to nonlinearly transform the input data to a high-dimensional space. The inner product process is performed by using the kernels function which is as follows (Chong and Zak, 2001):

$$K(x, y) = \sum_{j=0}^{p} \varphi_j(x)\varphi_j(y) = \varphi\varphi^T \quad (5)$$

K (x, y) is equivalent to $\varphi\varphi^T$ and is used to simplify complex calculations when calculating the values of (w, b) because the calculation of the conversion function ($\Phi$) for all vectors in the input space and then the inner product $\varphi\varphi^T$ procedure needs large calculations while the kernel function avoids calculating $\varphi(x)$ explicitly. This can be illustrated by the following equations:

$$f(x, w, b) = \text{sign}(w \cdot \varphi(x) + b) \quad (6)$$

$$w \cdot \varphi(x) = \sum_{i=1}^{N} \alpha_i d_i \varphi(x_i) \cdot \varphi(x) \quad \text{s.t} \quad \alpha_i > 0 \quad (7)$$

Equation 7 requires calculations during the processing of entries (Louis *et al*., 2010). Also $\varphi(x)$ does not have to be represented or found during the test. It is clear from the above that the classification equation becomes the final form:

$$f(x, w, b) = \text{sign}(\sum_{i=1}^{N} \alpha_i d_i k(x_i, x) + b) \quad (8)$$

Kernel transformation can be performed either using linear, polynomial or Radial Bias Functions (RBF) (Moore, 2001).

**Soft-margin support vector machine:** Soft margin SVM allows some data points to be on the wrong side of the margin. The optimization problem of the linear separable problem is like nonlinear separable but need the addition of slack variables ($\xi_i$) (Cortes and Vapnik, 1995). For data points on the correct side of the margin $\xi_i = 0$ for data points inside the margin $0<\xi_i<1$ and for misclassified data points $\xi_i>1$. The optimization problem is given as follows:

$$\text{Minimize}_{w, b, \xi} \frac{1}{2} \| w \|^2 + C \sum_{i=}^{n} \xi_i^k \quad (9)$$

Subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i \quad i = 1, 2, ..., n \quad (10)$$

The constraint $y_i(w^T x_i + b) \geq 1 - \xi_i$ can be written $y_i f(x_i) \geq 1 - \xi_i$ more concisely as which $\xi_i \geq 0$ together with is equivalent to $\xi_i = \max(1 - y_i f(x_i), 0)$. Hence, Eq. 9 has become an expressing as follows:

$$\text{Minimize}_{w, b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} (y_i, f(x_i)) \quad (11)$$

where, $L(y_i, f(x_i))$ is called the loss term and C is tuning parameter.

**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 (Hosmer *et al*., 2003). 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 (Ozkale and Arycan, 2016):

$$p(Y = y_i) = \pi(x_i)^{y_i} (1 - \pi(x_i))^{1 - y_i} \quad (12)$$

$y_i = 0$ or 1. The probability ($\pi_i$) can be defined mathematically in terms of explanatory variables and the logistic function as in the following formula:

$$\pi(x_i) = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}} \quad (13)$$

Where:

$\beta$: = Vector of parameters

$X_i$ = $\{1, x_{i1}, x_{i2}, ..., x_{ip}\}$

Row vector of independent variables. In order to simplify notation, we use the quantity $\pi(x) = E(Y/x)$ 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:

$$\text{logit}(\pi(x_i)) = \text{In}\frac{\frac{e^{X_i\beta}}{1+e^{X_i\beta}}}{1-\frac{e^{X_i\beta}}{1+e^{X_i\beta}}} \quad (14)$$

$$\text{logit}(\pi(x_i)) = \text{In}\frac{\frac{e^{X_i\beta}}{1+e^{X_i\beta}}}{\frac{1}{1+e^{X_i\beta}}} \quad (15)$$

$$\text{logit}(\pi(x_i)) = \text{In}(e^{X_i\beta}) = X_i\beta = (Bo+\sum_{j=1}^{p} B_j X_{ij}) \quad (16)$$

Whereas:

$\beta_0, \beta_1, ..., \beta_p$ = Unknown parameters were estimated
$X_{ij}$ = Independent variables

**Performance evaluation:** In this study, we will discuss several methods of evaluating the performance of (LR and SVM).

**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 (Soderstrom and Leitner, 1997) (Table 1).

**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:

$$\text{Accuracy} = (TP+TN)/N$$

Where:

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

**Sensitivity:**

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

Table 1: Confusion matrix

| | Prediction | |
|---|---|---|
| Classification | Negative | Positive |
| **Observations** | | |
| Negative | True Negative (TN) | False Positive (FP) |
| Positive | False Negative (FN) | True Positive (TP) |

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.

**Specificity:**

$$\text{Specificity} = \frac{TN}{TN+FP}$$

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.

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

$$\text{Prevalence} = \frac{TP+FN}{N}$$

**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 and Lemeshow, 2000).

$$A_{ROC} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} = \frac{1}{PN}\int_0^N TP*dFP \quad (17)$$

**Cohen's kappa coefficient:** When two binary variables are attempts by two individuals to measure the same thing, you can use Cohen's kappa (often simply called kappa) as a measure of agreement between the two individuals (Allouche *et al*., 2006). To compute kappa, you first need to calculate the observed level of agreement:

$$P_0 = \frac{TP+TN}{N}$$

This value needs to be compared to the value that you would expect, if the two raters were totally independent:

$$P_e = \left(\frac{(TP+FP)}{N} * \frac{(TP+FN)}{N}\right) + \left(\frac{(TN+FN)}{N} * \frac{(TN+FP)}{N}\right)$$

Then, the value of kappa is defined as:

$$K = \frac{P_0 - P_e}{1 - P_e}$$

The explanations below illustrate kappa's values with the appropriate estimates for each explanation (Carletta, 1996):

- Poor agreement = <0.20
- Fair agreement = 0.20-0.40
- Moderate agreement = 0.40-0.60
- Good agreement = 0.60-0.80
- Very good agreement = 0.80-1.00

## RESULTS AND DISCUSSION

**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 studied samples consist of 85 absence CKD and 68 presence CKD patients, the age of patients ranged from 12-90 years with a mean±SD of 42.75±17.39 years, respectively, also studied consist of 83 (54%) males and 70 (46%) females (Table 2).

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 their classification 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.

**Performance evaluation of models applied:** After dividing 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)

Table 2: Description of the study variables

| Variable name | Coding of variable | Types of variables |
|---|---|---|
| Class | 1 presence of CKD, 0 absence of CKD | Nominal |
| Sex | 1 male, 2 female | Nominal |
| Age | NA* | Numerical |
| Smoking | 1 smoked 2 non smoked | Nominal |
| Urea | N/A | Numerical |
| Creatinine | N/A | Numerical |
| Calcium | N/A | Numerical |
| Phosphor-us | N/A | Numerical |
| Alkaline phosphate-as | N/A | Numerical |
| Glucose | N/A | Numerical |
| Albumin | N/A | Numerical |
| Total Bilirubin | N/A | Numerical |

*Not available

Table 3: Confusion matrix and statistics for training dataset of LR

| | Prediction | |
|---|---|---|
| Classification | 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 (%); kappa coefficient 78.32 (%)

Now, 0 mean absence CKD, 1 mean presence CKD. The content of the response is unchanged: 65 cases of absence class and 43 cases of presence class were detected. From Table 3 we can see that the logistic regression model has been able to properly classify 97 cases out of 108 available. Classification errors were therefore, only 11. As it is possible to verify, the model has accuracy (89.81) but also the sensitivity and the specificity are >80% (81.40 and 95.38%). 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 kappa coefficient has achieved a good agreement of the model used for the value of 78.32%. As well as, we also found that the prevalence of the disease in the community for this model of the training dataset is 39.81%.

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 >80% (84.21 and 80.77%) the kappa coefficient has achieved a good agreement of the model used for the value of 64.07%. 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 $(AUC)_{ROC}$. Area under the curve: 0.8249.

Table 4: Confusion matrix and statistics for testing dataset of LR

| | Prediction | |
|---|---|---|
| Classification | 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 (%); Kappa coefficient 64.07 (%)

Table 5: Confusion matrix and statistics for training dataset of SVM

| | Prediction | |
|---|---|---|
| Classification | Absence 0 | Presence 1 |
| **Observations** | | |
| Absence 0 | 65 | 8 |
| Presence 1 | 0 | 35 |

Model accuracy 92.59 (%); Model sensitivity 81.40 (%); Model specificity 100 (%); Prevalence 39.81 (%); Kappa coefficient 84.04 (%)

Table 6: Confusion matrix and statistics for testing dataset of SVM

| | Prediction | |
|---|---|---|
| Classification | Absence 0 | Presence 1 |
| **Observations** | | |
| Absence 0 | 24 | 2 |
| Presence 1 | 2 | 17 |

Model accuracy 91.11 (%); Model sensitivity 89.47 (%); Model specificity 92.31 (%); Prevalence 42.22 (%); Kappa coefficient 81.78 (%)

ROC is plotted between the sensitivity (y axis) and the specificity (x axis). From Fig. 1 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 support vector machine model of the training dataset.

Table 5 summarizes, on average, the classification accuracy (92.59%) for the training dataset. In addition, we found sensitivity and specificity model (81.40 and 100%), respectively, these values have increased in comparison to the logistic regression model. On the other hand, the prevalence of the disease based on training dataset of SVM is 39.81%. So, kappa coefficient has achieved a very good agreement of the model used for the value of 84.04%. It is better than previous methods. From the Table 6 where the testing dataset has been achieved the model has accuracy (91.11%) but alsothe sensitivity and the specificity are >89% (89.47% and 92.31%), respectively, these are better results compared to all the previous methods of testing dataset. In addition, the kappa coefficient has achieved very good agreement of the model used for the value of 81.78%. As well as, we also found that the prevalence of the disease of the testing dataset is 42.22%. On the other hand, Area Under the Curve $(AUC)_{ROC}$: 0.9089 shows in Fig. 2. ROC is plotted between the sensitivity and the specificity, from Fig. 2 shows area under curve value is 90.89% this is the highest value for previous model.

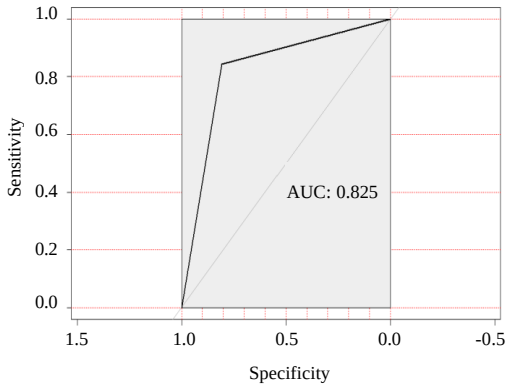


Fig. 1: (ROC) curve of testing dataset for LR

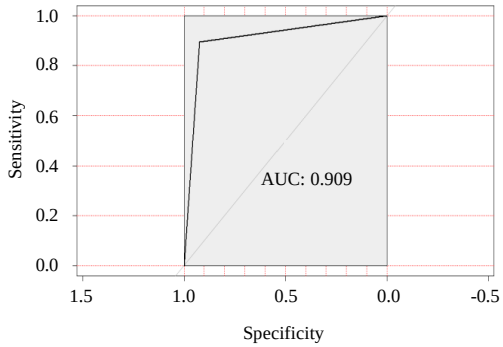


Fig. 2: (ROC) curve of testing dataset for SVM

**Comparison between models:** In this part of our study we discussed the classification using logistic regression model and support vector machine model. The two models were compared based on criteria (model accuracy, model sensitivity, model specificity, prevalence, kappa coefficient and area under curve (ROC)). The results showed that SVM Model was better than LR where the classification using SVM Model was more accurate and more efficient.

Table 7 shows model accuracy, model sensitivity, model specificity, prevalence, kappa coefficient and area under curve (ROC) for testing dataset. Because the best classifier based on testing dataset.

Table 7 summarizes, the accuracy of logistic regression was 82.22%. While the accuracy of support vector machine model was equal to 91.11%. This gives the preference for SVM 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 a SVM Model was equal to 89.47%. This gives preference to the SVM according to the model sensitivity criterion. In addition, the model

Table 7: Performance evaluation criteria between models

| Models | Logistic regression (%) | Support vector machine (%) |
|---|---|---|
| Model accuracy | 82.22 | 91.11 |
| Model sensitivity | 84.21 | 89.47 |
| Model specificity | 80.77 | 92.31 |
| Prevalence | 42.22 | 42.22 |
| (AUC) ROC | 82.49 | 90.89 |
| Kappa coefficient | 64.07 | 81.78 |

Table 8: Variables importance for SVM

ROC curve variable importance

| Variables | Importance |
|---|---|
| Creatinine | 100.000 |
| Urea | 95.997 |
| Albumin | 72.894 |
| Phosphorus | 68.641 |
| Calcium | 65.346 |
| Glucose | 40.659 |
| Alk.phosphatas | 33.862 |
| Age | 26.772 |
| Smoking | 3.753 |
| Sex | 0.000 |
| Total..Bilirubin | 0.000 |

specificity criterion for LR was 80.77%. The performance of the model was improved by using SVM. The value was 92.31%. 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 SVM Models was equal to 90.89%. The greater the value of area under curve (ROC). It was the best. In addition, the kappa coefficient was the best in SVM Model with a value of 81.78%, achieved a very good agreement comparison another models.

**Fitted final model and variables importance:** After analyzing and finding the optimal solution by the SVM Model, we had to know the most important variables (factors) affecting CKD patients in our study. By using VarImp function in the caret package for ROC curve variable importance in SVM Model. From Table 8 we found creatinine and urea factors were more affecting compared another independent variables for SVM Model.

Figure 3 shows significance of independent variables of SVM Model. 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. 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

Table 9: Variables importance for LR

GLM variable importance

| Variables | 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 |

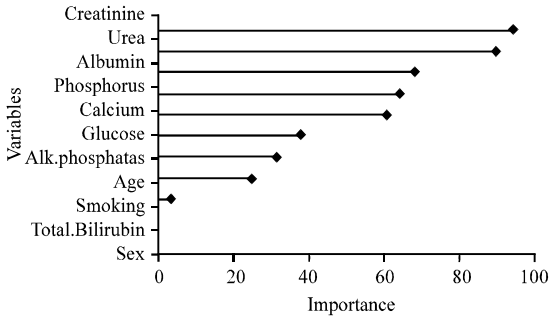


Fig. 3: Importance variables SVM Model plot

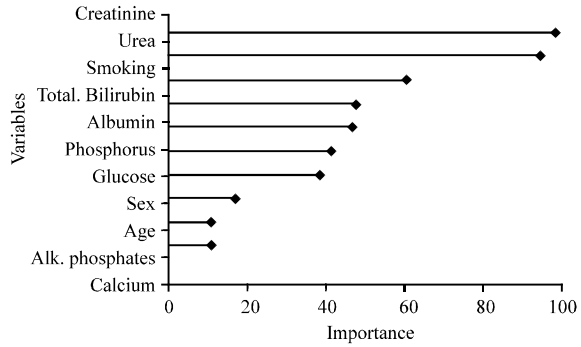


Fig. 4: Importance variables LR Model plot

variable urea and then smoking etc. Also, Fig. 4 shows significance of independent variables of logistic regression model. 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.

$$\log \text{odds} = (-10.40)+0.074x1+6.98x2$$

Where:

x1 = Urea independent variable

x2= Creatinine independent variable

Table 10: Factors affecting of CKD

| Variables | Estimate β | SE | Sig. | Wald values | df | p-values |
|---|---|---|---|---|---|---|
| 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 values

## CONCLUSION

In view of great importance of CKD and what may be caused of death and healthy crisis for community. In addition, being one of the diseases that have increased incidence in recent years, this research was achieved through the use of one of the traditional statistical models (logistic regression) with one of the models of the intelligent techniques to the following: the study has reached through comparison between of two classifier methods (logistic regression and support vector machine) in the classification of CKD patients relying on the blood test and based on evaluation criteria (model accuracy, model sensitivity, model specificity, kappa coefficient and area under curve (ROC)) that the method of SVM is the best methods used in this study.

Depending on model accuracy, the accuracy of support vector machine model was equal to 91.11%, also the accuracy of logistic regression was 82.22%. This gives the preference for SVM according to the accuracy of the model.

The kappa coefficient which measures the compatibility of the observed data in model with the expected data, showed a clear advantage of SVM method compared to LR where it was valued 81.78%, this means it achieved a very good agreement.

The results indicated that the factors had greatest effect on the data of patients with chronic renal failure using the two methods (LR and SVM) that the variables (creatinine and urea) are the most effective and significant variables. The prevalence of disease at community approximately 42%, it's a big percent. Calls for action to reduce prevalence among the population.

## REFERENCES


Allouche, O., A. Tsoar and R. Kadmon, 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the True Skill Statistic (TSS). J. Appl. Ecol., 43: 1223-1232.

Bhatla, N. and K. Jyoti, 2012. An analysis of heart disease prediction using different data mining techniques. Intl. J. Eng. Res. Technol., 1: 1-4.

Boser, B.E., I.M. Guyon and V.N. Vapnik, 1992. A training algorithm for optimal margin classifiers. Proceedings of the 5th Annual Workshop on Computational Learning Theory, July 27-29, 1992, Pittsburgh, Pennsylvania, USA., pp: 144-152.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining Knowl. Discov., 2: 121-167.

Byvatov, E., U. Fechner, J. Sadowski and G. Schneider, 2003. Comparison of support vector machine and artificial neural network systems for drug nondrug classification. J. Chem. Inf. Comput. Sci., 43: 1882-1889.

Carletta, J., 1996. Assessing agreement on classification tasks: The kappa statistic. Comput. Ling., 22: 249-254.

Chen, S.T., Y.H. Hsiao, Y.L. Huang, S.J. Kuo and H.S. Tseng *et al.*, 2009. 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. Korean J. Radiol., 10: 464-471.

Cho, S.B. and H.H. Won, 2003. Machine learning in DNA microarray analysis for cancer classification. Proceedings of the 1st Asia-Pacific Bioinformatics Conference, February 4-7, 2003, Adelaide, Australia.

Chong, E.K. and S.H. Zak, 2001. An Introduction to Optimization. 2nd Edn., John Wiley & Sons, Hoboken, New Jersey, USA., ISBN:9780471391265, Pages: 496.

Colas, F. and P. Brazdil, 2006. Comparison of SVM and some older classification algorithms in text classification tasks. Proceedings of the 2006 IFIP International Conference on Artificial Intelligence in Theory and Practice, August 21-24, 2006, Springer, Boston, Massachusetts, ISBN:978-0-387-34654-0, pp: 169-178.

Cortes, C. and V. Vapnik, 1995. Support-vector networks. Mach. Learn., 20: 273-297.

Cristianini, N. and J. Shawe-Taylor, 2000. An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK.

George, Y.M., H.H. Zayed, M.I. Roushdy and B.M. Elbagoury, 2014. Remote computer-aided breast cancer detection and diagnosis system based on cytological images. IEEE. Syst. J., 8: 949-964.

Hosmer Jr, D.W., S. Lemeshow and R.X. Sturdivant, 2003. Applied Logistic Regression. 3rd Edn., John Wiley & Sons, Hoboken, New Jersey, USA., ISBN:978-0-470-58247-3, Pages: 479.

Hosmer, W.D. and S. Lemeshow, 2000. Applied Logistic Regression. 2nd Edn., John Wiley and Sons, New York, USA., ISBN-10: 0471356328, Pages: 392.

Inan, D. and B.E. Erdogan, 2013. Liu-type logistic estimator. Commun. Stat. Simul. Comput., 42: 1578-1586.

Jha, V., G. Garcia-Garcia, K. Iseki, Z. Li and S. Naicker *et al*., 2013. Chronic kidney disease: Global dimension and perspectives. Lancet, 382: 260-272.

Louis, B., V.K. Agrawal and P.V. Khadikar, 2010. Prediction of intrinsic solubility of generic drugs using MLR, ANN and SVM analyses. Eur. J. Med. Chem., 45: 4018-4025.

Moore, A.W., 2001. Support vector machines. Master Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Ozkale, M.R. and E. Arýcan, 2016. A new biased estimator in logistic regression model. Stat., 50: 233-253.

Sarwar, A. and V. Sharma, 2014. Comparative analysis of machine learning techniques in prognosis of type II diabetes. AI. Soc., 29: 123-129.

Soderstrom, I. R., and D.W. Leitner, 1997. The effects of base rate, selection ratio, sample size and reliability of predictors on predictive efficiency indices associated with logistic regression models. Proceedings of the Annual Meeting on the Mid-Western Educational Research Association, October 15-18, 1997, Chicago, Illinois, pp: 1-25.

Tan, P.N., M. Steinbach and V. Kumar, 2005. Introduction to Data Mining. 1st Edn., Pearson Addison Wesley, New York, ISBN-13: 978-0321321367, Pages: 769.

Vijayarani, S. and S. Dhayanand, 2015. Kidney disease prediction using SVM and ANN algorithms. Int. J. Comput. Bus. Res., 6: 1-12.