IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 21, Issue 1, Ver. I (Jan - Feb 2019), PP 86-92
www.iosrjournals.org

Cardiac Catheterization Procedure Prediction Using Machine Learning and Data Mining Techniques

Huda Kutrani[1], Saria Eltalhi [2]
[1](Health Informatics Department, Faculty of Public Health, University of Benghazi, Libya)
[2](Computer Science Department, Faculty of Education, University of Benghazi, Libya)
Corresponding Author: Huda Kutrani,

Abstract: *Although catheterization is an important tool in the diagnosis and the treatment of cardiovascular diseases, it may cause different complications such as death or myocardial infarction during diagnosis. Data mining techniques are used for the construction of a cardiac catheterization Prediction System (CCPS) for whom catheterization is needed; therefore, it can decrease the complications of Cardiac catheterization Procedure. The aim of this study is to predict whether a patient needs a cardiac catheterization procedure or not. WEKA software was used in this experimental evaluation study of the Home dataset. Five classification algorithms were used for the prediction of catheterization procedure based on the prediction Accuracy, True Positive, True Negative, and ROC area. This study concluded that J48 without smoker attribute was a well-suited model for the prediction of whether a patient needs a cardiac catheterization procedure or not.*

Keywords: *Data mining, supervised Classification Algorithms, Machine Learning, Cardiac catheterization.*

Date of Submission: 12-01-2019 Date of acceptance: 27-01-2019

I. Introduction

There are one billion adults living with cardiovascular diseases (CVD) worldwide[1]. Although cardiac catheterization is an important procedure to diagnose and treat cardiovascular diseases, it can cause some complications such as death and myocardial infarction [1,2]. More than one million patients undergo cardiac catheterization procedure in the USA yearly [1]. Therefore, the cardiac catheterization procedure can increase health care expenditures and be a heavy burden on the economies of countries, especially developing countries [1,3]. Although a huge amount of cardiovascular disease data was collected, physicians cannot always make the correct decision for all patients in need of the cardiac catheterization procedure. Boyles (2010) reported that in the USA, about 20% of patient who did cardiac catheterization procedure, they do not need it [1]. According to Benghazi Heart Disease Center, about 18% of patients who did cardiac catheterization procedure, they do not need it.

Clinical data is increasing, to analyze, these huge data without using a computer-based analysis system is a very difficult task. The computer-based analysis system has many mechanisms to support the medical practitioner to make a good decision in treatment and diagnosis [4]. Data mining is the most useful technique to discover hidden information and knowledge of the huge amount of patient data sets. This technique can discover useful patterns. The discovered patterns must be meaningful to lead to some advantage in predicting and making the right decision. Therefore, these patterns are easy to understand and can be used to make certain decisions for the development of healthcare [5,6,7].

This study aims to find the most efficient data mining procedure to use for predicting whether a patient needs a cardiac catheterization procedure or not. In this study, the classification task is employed to predict the need for cardiac catheterization procedure; based on the Prediction Accuracy, True Positive and True Negative by different Machine Learning algorithms classification.

Related Work

Alizadehsani et al. [8] have evaluated different classification techniques; three classification algorithms were employed to analyze the data set, SMO, Naïve Bayes, and a proposed ensemble algorithm. The data set consisted of 303 patients. The implementation of the classification methods was done by java, on top of Weka. The study showed that the proposed ensemble method had the highest accuracy rate of all methods used. The Naïve Bayes and SMO methods had nearly the same accuracies.

Boshra Baharami et al.[9]] have evaluated and compared different classification techniques for heart disease diagnosis. Classification techniques such as J48 Decision tree, k-Nearest Neighbors (kNN), Naïve Bayes (NB) and SMO were used. WEKA software is used for implementing the classification algorithms. The data

set contained 13 medical attributes. Models were evaluated using six different performance measures. The study showed that j48 decision tree achieved the highest value in accuracy.

Nidhi Bhatla et al. [10] Analysed diverse data mining techniques for heart disease prediction, namely, Naive Bayes, Decision List, and KNN; for efficient diagnosis of heart disease. The research worked on incorporating two or more attributes, i.e. obesity and smoking. Decision tree had outperformed any other techniques regarding accuracy; with the help of a genetic algorithm and feature subset selection using WEKA 3.6.6. Data mining techniques were applied to 15 attributes. The analysis showed that the Neural Network and Decision Tree had high accuracy (100% and 99.62% respectively).

II. Material And Methods

Proposed model

The proposed architecture of a cardiac catheterization prediction system shown in Figure 1. It consists of a training dataset and a testing dataset which were extracted from a Home data set. WEKA data mining tool was used to implement the cardiac catheterization prediction system. Different five supervised machine learning classification methods were used to evaluate the prediction system.

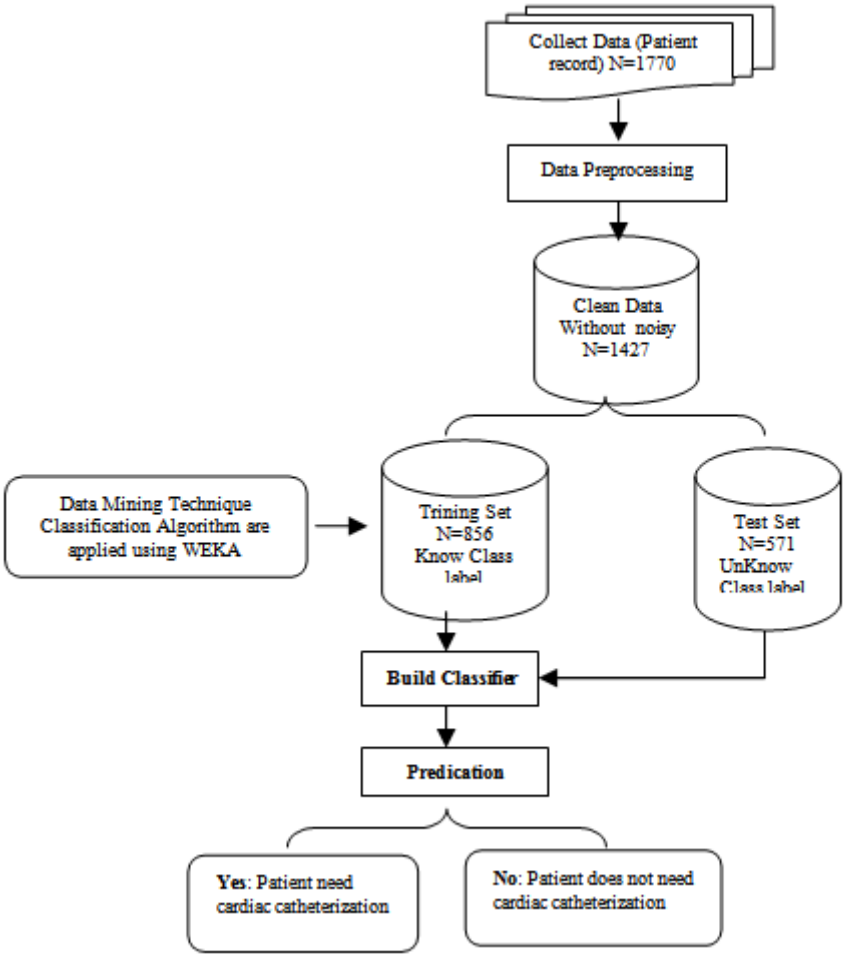

Figure (1): Proposed model

Home data set

This study was conducted by using Home data set from Benghazi Heart Disease Center. The dataset is real data of actual patients who underwent cardiac catheterization in the center from December 2003 to May 2007. The full number of patients' records were 1770 and 11 attributes. After preprocessing data, the Home data set was 1427 instances with 9 attributes including class attribute.

Data preprocessing

Data preprocessing is an important step of many that influences the performance and quality of prediction of machine learning models; it included:

Missing values: There were three to five attributes that have missing values in 230 patients' records; these records were removed from the Home data set.

Non-suitable records: The study aims to predict whether a patient needs a cardiac catheterization procedure or not among adults with cardiovascular diseases. Thus, 113 recodes were removed from the Home data set; because of age (under 15) and a congenital heart defect.
Choosing the attributes: an assessment of the 11 attributes was conducted and attributes that were expected not to contribute to the results such as patients' address were removed.

The Final Home data set: It was 1427 instances with 9 attributes including class attribute.

Description of the Home data Set

This Home data set consists of 1427 instances with 9 attributes including class attribute. The class attribute was classified as two classes of *'Yes'* and *'No'* in terms of the need for cardiac catheterization. 77.2% of the Home data set represented the 'Yes' classification, and 'No' classification was 22.8% of the Home data set. The mean age was 58 years; the median was 58, males were 58.3% and 41.7% were females. Table 1 shows a more detailed attribute description.

Table (1): Data Set attributes

Attribute	Representation	Type	Domain
Gender	Patient's gender	Nominal	{Male,Female}
Age	Age in years	Numerical	Age > 20 years
MI	Myocardial Infarction	Nominal	{Positive, Negative}
HTN	Hypertension	Nominal	{ Positive, Negative}
DM	Diabetes	Nominal	{ Positive, Negative}
Angina	Chest pain	Nominal	{ Positive, Negative}
Smoker	Smoker patient	Nominal	{ Positive, Negative}
IHD	Ischemic heart disease	Nominal	{ Positive, Negative}
Class	Class	Nominal	{Yes, No}

Training and Testing data set: The Home data set consists of 1427 instances. It was divided into a Training data set with 60% of the Home data set and a Testing data set with 40% of the Home data set.

Machine learning classification models

The machine learning model learns from past input data to make a future prediction as output. Machine Learning has the ability to learn and improve its performance without human instructions. The algorithms get adjusted based on the training sample and an error signal [4,6]. Machine Learning algorithms automatically test and analyze all prediction variables to prevent overlooking any potentially important predictor variables even if it was unexpected. Therefore, ML is a powerful tool [5,6,7]. The machine learning techniques used in analyzing the medical data including *Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), J48 and Naive Bayes.*

Support Vector Machine (SVM): It is a supervising learning method that looks at data and sorts it into one of two categories. The SVM takes an input and predicts where to classify the classes optimally by creating the approach between two data clusters. This algorithm attains its high accuracy using nonlinear features called kernels [4,10].

J48: J48 decision tree is a supervised classification method. It uses the Divide and Conquer approach to classify data. J48 algorithm can reduce the complexity by dividing the large dataset into a small dataset; thus improving the performance of classification [11].

K-Nearest Neighbors (KNN): it is one of the lazy learning algorithms used for classification. It uses the "K" nearest neighbors where there is no training data. It considers the neighbors around an object; its role is to assign a class of the most common among its k nearest neighbors (k is small and a positive integer.) [12,13].

Artificial Neural Network (ANN): is a model of reasoning and processing information based on the human brain. It is made up of many layers (arranged in layers), where the output of each layer is an input for the next layer, without feeding back the previous layer. The data is received in the input layer then transmitted to a hidden layer where the data processing is done; the results of the processing are sent to the output layer [12].

Naive Bayes (NB): It is used for classification based on Bayes' theorem. It is a statistical classifier, and it is most commonly used in machine learning [14].

WEKA

Waikato Environment for Knowledge Analysis (WEKA) was developed by the University of Waikato in New Zealand [14,15]. It is an open-source software system where the code is publically available, and it has some machine learning algorithms (ML) can be used for data mining tasks [14,16].

Data set file in WEKA: WEKA requires a file with the format (arff). The original home cardiac catheterization data set file stored in the Microsoft Excel (spreadsheet) with the format (xls). This study used WordPad software to convert the Home data set into (arff) format.

Experiment

The models gained learning from the Training data set. In this study, five Machine Learning algorithms were applied to the Test data set. In order to classify and predict which patients need the cardiac catheterization. The five Machine Learning algorithms were Naïve Bayes, Neural Network, J48, SVM, and KNN algorithms. The Test data set was tested five times in WEKA, once for each classifier. The evaluation of Machine Learning models was based on the prediction Accuracy, True Positive, ROC area, True Negative, Kappa and Mean absolute error. In this study, attributes were tested using logistic regression to identify the significant factors that affect the prediction model. Age, gender, and smoker attributes had negative influences on the accuracy and the performance of the model. The Test data set was tested without age, gender, or smoker attributes using five Machine Learning algorithms.

Statistical analysis

The outcome of binary classifier as shown in the Table 2.

Table (2): different outcomes of a two-class prediction

Actual class	**Predicted class**	
	YES	**NO**
YES	True Positive **(TP)**	False Negative (FN)
NO	False Positive (FP)	True Negative **(TN)**

- ***True positive (TP):*** The patients have been correctly predicted as positive (predicted need cardiac catheterization and they do need cardiac catheterization.).
- ***True negative (TN):*** The patients have been correctly predicted as negative (predicted not needing cardiac catheterization and they do not need cardiac catheterization.)***When TP and TN close to 100%, the model becomes a perfect model.***
- ***Correctly Classified Instances (CCI):*** It represents the percentage of patients that have been correctly diagnosed, both who need and not need cardiac catheterization. Also known as Accuracy Correct Classified (ACC) [14].
- ***Kappa*****:** *"Kappa measures the agreement of prediction with the true class. It calculates the difference between the predictions with the observed agreement with that expected by chance. The kappa statistic value is a value between 0-1. A value greater than 0 means that the classifier is better than chance"*[15:16].
- ***ROC Area*****:** *"ROC Area is the area under the ROC curve. ROC curve shows how well a classifier is at distinguishing between positive and negative instances. ROC area can be used to evaluate the quality of the classifier and its ability to separate positive and negative instances. ROC Area can be a value between 0.5 and 1, 1 being an optimal classifier and 0.5 being comparable with a chance"*[15:16]. Therefore, a good model has a ROC area above 0.6.
- ***Mean absolute error (MAE):*** A comparison between predictions and the eventual outcomes. It can be calculated by 1- ACC. A good model has a very small mean absolute error [14].

III. Result & Discussion

Table 3 showed a comparison between the five different Machine Learning algorithms. In the term Correctly Classified Instances; Naïve Bayes, Neural Network and J48 algorithms had the highest percentage of patients who have been correctly diagnosed. This means that about 87% of patients had the correct positions in the classification pattern (TP: patients need a cardiac catheterization procedure and TN: patients do not need a cardiac catheterization procedure.). Also, SVM and KNN algorithms had a good percentage in the term Correctly Classified Instances with 86% and 85% respectively. These results agree with Kodati & Vivekananda study which was a comparison of data mining classification techniques. The purpose was to analyze heart disease datasets obtained from the Cleveland dataset using WEKA program [14]. Moreover, Naïve Bayes,

Neural Network and J48 algorithms had a good value of TP. This means the percentage of the patients have been correctly diagnosed for needing a cardiac catheterization procedure.

Although Naïve Bayes, Neural Network and J48 algorithms had a good ROC area values which are considered good classifiers' performance, the percentages of True Negative (TN) were not good enough (53%, 65%, and 61% respectively). When compared the percentages of TN among the five algorithms; SVM, Neural Network and J48 algorithms had the highest percentage; as shown in Table 3. These algorithms were the most suitable models to achieve the aim of this study. Moreover, they performed well in predictive studies in the field of heart disease [14,15,17,18]. In this study SVM, Neural Network and J48 algorithms were selected to improve their accuracy and performance.

Table (3): Comparative analysis of the prediction results among algorithms

Algorithms	Correctly classified instances % (Accuracy)	TP %	TN %	ROC area	Kappa	Mean absolute error
Naïve Bayes	87.5	97	53	0.94	0.6	0.17
SVM	86	91	66	0.8	0.6	0.14
KNN	85	93	55	0.83	0.5	0.15
J48	87.2	94	61	0.94	0.6	0.15
Neural network	87.4	94	65	0.94	0.6	0.15

Many studies suggested removing attributes to improve the accuracy and performance of a model [15,19,20,21]. In this study, attributes were tested using logistic regression to identify the significant factors that affect the prediction model. Age, gender, and smoker attributes had negative influences on the accuracy and the performance of the model.

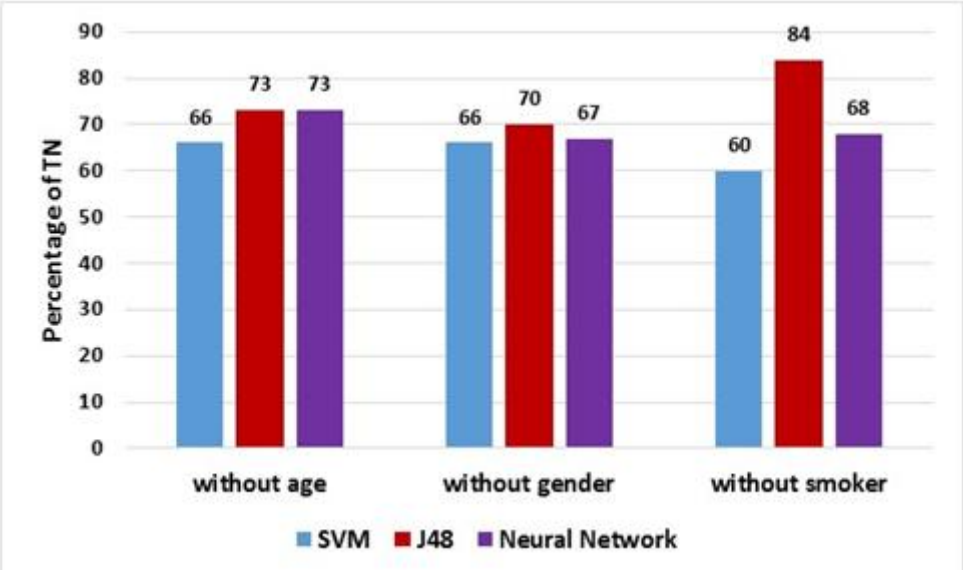

Figure (2): showed results of TN percentage according to three best algorithms in this study

Figure 2 shows the results of TN describing the percentage of patients not needing a cardiac catheterization procedure by the chosen algorithms. The home dataset without smoker using J48 algorithm provided a higher percentage of TN (84%). Followed by the home dataset without age attribute with a good TN percentage using J48 and Neural Network (73% for both). In a Swedish study (2015); they found that the home dataset without gender provided a higher accuracy model than the home dataset without age [15]. Despite the prevalence of tobacco use among Libyan adult males is 66% according to WHO's report (2017)[22]. The home dataset showed only 20.3 % of adult males smoke with age mean of 58 years and median was 58 years.

Table (4): Comparative analysis of the prediction results from J48 algorithm

Algorithm	Correctly classified instances %	TP %	TN %	ROC area	Kappa	Mean absolute error
J48	87.2	94	61	0.94	0.6	0.15
J48 without age	87.6	92	73	0.91	0.63	0.18
J48 without gender	86.5	91	70	0.92	0.6	0.16
J48 without smoker	89	91	84	0.94	0.7	0.15

Figure 2 shows that the J48 algorithm was the best of the prediction algorithms. Table 4 shows the comparative analysis of the predicted results of the J48 models. The J48 without smoker was the best model to predict whether a patient needs a cardiac catheterization procedure or not. It had the highest Correctly Classified Instances (89%). The ROC area value was 0.94 close to the perfect classifier represented by 1. Moreover, the True Negative was 84%, which was the highest value in all algorithm models, and the True Positive was 91%.

Even though, the accuracy of the J48 without smoker model was 89%, and False True diagnosis and True Positive diagnosis were 84% and 91% respectively, they are not accurate enough to actually be implemented in medical care. In the home data set, 325 of 1427 patients did a not needed cardiac catheterization procedure. The J48 without smoker could have saved 273 patients of 325. Therefore, the experiment should be repeated with the bigger home dataset without the smoker variable to evaluate whether it is possible to achieve accuracy, True Negative diagnosis and True Positive diagnosis of around 95%.

IV. Conclusion

Data mining and machine learning techniques do assist in finding the hidden information and knowledge in data that predict diagnosis and treatment of diseases. Classification is one of these techniques. This study has implemented five data mining and machine learning techniques by using the Home dataset to predict whether a patient needs a cardiac catheterization procedure or not. Comparative analysis using WEKA software; it found that Naïve Bayes, Neural Network and J48 algorithms had the highest accuracy. However, J48 without smoker attribute was the best model to predict whether a patient needs a cardiac catheterization procedure or not with an accuracy of 89%.

Acknowledgement

The authors appreciate the contribution of Benghazi Heart Disease Center as a source of the Home data set. A special thanks to statistics unit staff for their help.

References

[1]. World Health Organization. Cardiovascular diseases (CVDs). https://www.who.int/cardiovascular_diseases/en/. [Accessed 3rd January 2019].
[2]. K. T. Keerthana. "Heart Disease Prediction System using Data Mining Method". International Journal of Engineering Trends and Technology (IJETT), Vol. 47, Issue 6, 2017, pp. 361-63.
[3]. G.W. Reed, M.L. Tushman, and S.R. Kapadia. "Effective Operational Management in the Cardiac Catheterization Laboratory". Journal of the American College of Cardiology, Vol. 72, Issue 20, 2018, pp. 2507-17.
[4]. B. Srinivasan and K. Pavya. "A STUDY ON DATA MINING PREDICTION TECHNIQUES IN HEALTHCARE SECTOR". International Research Journal of Engineering and Technology (IRJET). Vol. 3, Issue 3, 2016, pp. 552-56.
[5]. M. Gera and S. Goel. "Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity". International Journal of Computer Applications, Vol. 113, Issue 18, 2015, pp. 22-29.
[6]. R. Manimaran and M. Vanitha. "Prediction of Diabetes Disease Using Classification Data Mining Techniques". International Journal of Engineering, Vol. 9, Issue 5, 2017, pp. 3610-14.
[7]. M. Ramageri. "Data mining Techniques and Applications". Computer Science and Engineering, Vol. 1, Issue 4, 2010, pp. 301-05.
[8]. R. Alizadehsani. "Diagnosis of Coronary Artery Disease Using Data Mining Techniques Based on Symptoms and ECG Features". European Journal of Scientific Research, Vol. 82, Issue 4, 2012, pp. 542-53.
[9]. B. Bahrami, and M.H. Shirvani. "Prediction and Diagnosis of Heart Disease by Data Mining Techniques". Journal of Multidisciplinary Engineering Science and Technology (JMEST), Vol. 2, Issue 2, 2015, pp. 3140-59.
[10]. N. Bhatla, and K. Jyoti. "An Analysis of Heart Disease Prediction using Different Data Mining Techniques". International Journal of Engineering Research & Technology (IJERT), Vol. 1, Issue 8, 2012, pp. 2274-78.
[11]. L. Jena and N.K. Kamila. "Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease". Emerging Research in Management & Technology, Vol. 4, Issue 11, 2015, pp. 110-18.
[12]. W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis". MDPI, Vol. 2, Issue 2, 2018, pp. 1-17.
[13]. D.P. Varghese and P.B. Tintu. "A survey on health data using data mining techniques," Engineering and Technology, Vol. 2, Issue 7, 2015, pp. 713-20.
[14]. S. Kodati and R. Vivekanandam, "Analysis of Heart Disease using in Data Mining Tools Orange and Weka". Vol. 18, Issue 1, 2018, pp. 16-21.
[15]. A. Olsson and D. Nordlöf. "Early screening diagnostic aid for heart disease using data mining: an evaluation using patient data that can be obtained without medical Equipment". Degree Project in Computer Science Thesis. KTH Royal Institute of Technology; Sweden, 2015.
[16]. S. Sakr, R. Elshawi, A. Ahmed, W.T. Qureshi, C. Brawner, S. Keteyian, M.J. Blaha, M.H. Al-mallah, and H. Ford. "Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) Project". PLoS ONE, Vol. 13, Issue 4, pp. 1-18. e0195344. https://doi.org/10.1371/journal.pone.0195344Vascular, 2018, pp. 1-18.

[17]. V S Kumbhar, K S Oza, and R R Mudholkar. "Classification of Heart Data using Weka". 2nd International Conference on Cognitive Knowledge Engineering, Marathwada University, Aurangabad. 2016, pp. 1-3.
[18]. S.K. Mandal, A. Gupta, A. Mukherjee, and A. Mukherjee. "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review". Advances in Computational Sciences and Technology, Vol. 10, Issue 7, 2017, pp. 2137-59.
[19]. H.N. Mufti. "A Data Mining Approach for Predicting Delirium After Cardiac Surgery". M.Sc. Thesis, Dalhousie University, 2014.
[20]. R. Assari, P. Azimi, and M.R. Taghva. "Heart Disease Diagnosis Using Data Mining Techniques". International Journal of Economics & Management Sciences, Vol. 6, Issue 3, 2017, pp. 1-5.
[21]. S.B. Patil and D.Y. Kumaraswamy. "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction". IJCSNS International Journal of Computer Science and Network Security, Vol. 9, Issue 2, 2009, pp. 228-235.
[22]. World Health Organization. "WHO report on the global tobacco epidemic, 2017: Country profile-Libya" WHO Framework Convention on Tobacco Control (WHO FCTC) status. Report 2017.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Huda Kutrani and Saria Eltalhi. "Cardiac Catheterization Procedure Prediction Using Machine Learning and Data Mining Techniques" IOSR Journal of Computer Engineering (IOSR-JCE) 21.1 (2019): 86-92